UNDERSTANDING PROTEIN FOLDING AND ASSEMBLY AT THE OUTER MEMBRANE OF PATHOGENIC GRAM-NEGATIVE BACTERIA

A Dissertation
Presented to
The Academic Faculty

By

Anthony J. Hazel

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Physics

Georgia Institute of Technology

December 2018

UNDERSTANDING PROTEIN FOLDING AND ASSEMBLY AT THE OUTER MEMBRANE OF PATHOGENIC GRAM-NEGATIVE BACTERIA

Approved by:

Dr. James C. Gumbart, Advisor
School of Physics
Georgia Institute of Technology

Dr. Chris Chipot
Laboratoire International Associè CNRS-UIUC
Universitè de Lorraine

Dr. Raquel Lieberman
School of Chemistry and Biochemistry
Georgia Institute of Technology

Dr. Harold Kim
School of Physics
Georgia Institute of Technology

Dr. Kurt Wiesenfeld
School of Physics
Georgia Institute of Technology

Date Approved: December 10, 2018

A great quote to start the thesis

George P. Burdell

A great dedication goes here.

ACKNOWLEDGEMENTS

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Our understanding of the microscopic world inside of our own bodies has lead to astronomical advances in medicine over the last century. As we've moved from the 20th to the 21st century, however, our microscopes have become increasingly digitized. We can now observe inside own cells through the 0's and 1's of our computers through the power of molecular dynamics (MD) simulations. Molecular dynamics (MD) has become an invaluable tool for microbiologists, as it provides infinitely reproducible, 100% sterile laboratories for experimentation on the smallest components of life: individual atoms. The late Klaus Schulten used to refer to MD simulations as a computational microscopic, and with current hardware, the scope of our lenses as increased dramatically, enabling us to see entire organelles and even viruses in perfect detail [1, 2]. Special purpose machines such as Anton and Anton 2 now let us run this microscopic for extraordinary lengths of time, computationally speaking [3, 4]. For this thesis, I will discuss one of the many common ways to further enhance MD simulations that not only help us observe but also quantify a wide variety biological and molecular phenomena.

Protein folding is one of the great unsolved problems of molecular biology. To this day, we are still learning more and more about how even small structural motifs within a larger protein fold, and the folding pathways of entire proteins that are more than a couple dozen residues are still largely unknown or unexplored. In this thesis I will discuss the technique of umbrella sampling (US) with MD simulations and utilize it to study the folding free energy landscapes of two simple, model peptides: Ala_{10}, used to model α-helix folding, and GB1, used to model β-sheet folding. These simple peptides can produces surprisingly complex and nuanced folding pathways. I performed some of the first assessments of the widely used CHARMM force fields in their ability to accurately recover experimental folding properties of these two peptides, and found that while they work quite well for α-helices, they are far less accurate for β-sheets. I also observed the important role that water

plays in the folding process, particularly in its ability to stabilize unfolded conformations where the backbone is more exposed to the surrounding solvent. A delicate balance of hydrogen bonding strength is needed between the water and the peptide and the peptide with itself. This was the most apparent with the Drude polarizable force field, where hydrogen bonding interactions with the backbone amide groups were too weak, both with water and with the backbone carbonyl groups. While this helped to more accurately describe the unfolded conformations compared to its fixed charged brethren, it lead to a significant destabilization of the folded state. By fine tuning the amide polarizability, I was able to recreate the correct folded fractions as expected from experiment.

I my applied my results for the GB1 β-hairpin peptide to a much larger β-sheet structure, the β-helical passenger domains of autotransporter proteins. These proteins are used by pathogenic Gram-negative bacteria to secrete virulence factors out of the cell. It was originally thought that the folding of this unique yet ubiquitous structure was the energy source for this transport process, given the dirth of traditional energy sources at the outer membrane. Although this theory has questioned, the environment at the outer membrane along with physio-chemical properties of these β-helices suggest that they fold along a vectorial pathway *in vivo* [5], much faster than the concerted pathway they take *in vitro* [6, 7]. Folding free energy landscapes calculated from self-learning adaptive umbrella sampling (SLUS) [8] of multiple bands from both termini of the pertactin passenger domain, a common model for used to study virulence factor secretion, reveal the overall lack of stability throughout the helix when no other neighboring structures are present. Folding one neighboring band, however, leads to downhill, cooperative folding events which propagate through the helix, with the two terminal bands acting as the scaffolds. These results suggest how the cell utilizes both the instability and cooperativity to quickly fold the helix only during secretion, while preventing from unfolding otherwise.

Lastly, I also applied SLUS to another important biological question, that of protein assembly. The AcrAB-TolC multidrug efflux pump in *Escherichia coli* is way of its main

sources of antibiotic resistance, where it ejects antimicrobial drugs and other cytotoxic agents from the cell before they reach their intended target. Conformational free energy landscapes of the membrane fusion protein, AcrA, which is responsible for assembly of this complex, reveals two main conformations, one of which is assembly compatible. Mutational analysis and equilibrium simulations show that residue-residue contacts at an interfacial site between two of its domains that are known to interact with the inner membrane-bound transporter, AcrB, disrupt pump assembly and function by affecting AcrA's flexibility. Free energy landscapes of AcrB-bound AcrA also reveal a conformational selection binding mechanism in one of AcrB's two binding sites, indicating that the two binding sites may serve to stimulate disparate processes, one at the inner membrane and the other at the outer membrane.

CHAPTER 1

MOTIVATION

From the germ theory of disease and the use of antiseptics to the discovery of penicillin and the development of the first polio vaccine, the extension of man's average lifespan due to the evolution of modern medicine far outpaces any advantages gained through natural evolution over the same time span. Biomedical research is such an important factor in today's society that, outside of the Department of Defense, the National Institutes of Health accounts for the largest share of federal funding for basic research in the US [9]. This also drives a $333 billion dollar pharmaceutical industry for developing, testing, and marketing new drugs [10]. While many drugs on the market today were originally derived or extracted from naturally occurring sources, more and more drugs are being designed to fulfill a specific purpose. As most drug trials proceed by testing thousands upon thousands of compounds, performing such tests computationally could drastically reduce the cost of drug research by reducing the initial pool of compounds. Because of the economic advantages, much work has been done in the late 20th and early 21st centuries to improve upon current design schemes or to develop new ones.

Drugs work by targeting specific processes within the cell. This often involves enhancing or suppressing the activity of a protein, whether in the cells of the patient or the cells of foreign invaders. In such cases, the drug molecules will bind to the protein of interest, thus affecting its activity in the desired way. How do pharmacologists design drugs for specific proteins? First, the structure of the protein is determined, typically by X-ray crystallography. From this structure, one can predict where a drug molecule is likely to bind, called a binding pocket. A computational procedure called "docking" determines how well a drug molecule can bind in the binding pocket [11]. Libraries of thousands of drug molecules can be reduced down to a hundred or so, effectively eliminating the first round of drug testing.

Docking, however, does have its limitations. Proteins are inherently dynamical systems that exist within a highly complex environment. There is no guarantee that the crystal structure represents a biologically accessible or relevant state. Also, new potential drug binding sites could arise as the structure of the protein fluctuates that would not have been observable with a single static image. Drug designers need a way to observe the dynamics of the proteins to resolve these issues. Luckily, a well-established technique exists for such a purpose: molecular dynamics (MD). MD approximates molecular motions using a classical ball-and-spring model and has become a standard computational tool for molecular biologists, biophysicists, and pharmacologists. Although unable to describe certain quantum mechanical phenomena known to be present in molecular systems, MD has proven to be an efficient and accurate means of observing and quantifying biological systems.

While at first only capable of simulating small proteins for short amounts of time (think hundreds of atoms for a couple of picoseconds), the scale of current MD simulations is astounding, and the phenomena observed are quite extraordinary. The NIH Center for Macromolecular Modeling, for example, is constantly pushing the boundaries of MD, producing some of the largest systems ever simulated, including the entire viral capsids of HBV and HIV-1 ($\sim$10 million atoms) [1] and the low-light adapted chromatophore of *Rhodobacter sphaeroides* ($\sim$100 million atoms) [2]. Special purpose machines, such as Anton and Anton II from the D. E. Shaw Research group, also allow us to simulate biological systems on the order of milliseconds [3]. Drug binding and unbinding can take on the order of minutes for certain ligand-protein complexes [12]; so while MD may still be just scratching the surface of biologically relevant timescales, techniques exist to "speed up" the simulations. D. E. Shaw and colleagues recently demonstrated one such technique – temperature accelerated molecular dynamics (TAMD) – to study how a ligand dissociates from a human adenosine receptor [4]. They also identified mutants that did not affect ligand-binding affinities but drastically affected ligand-dissociation kinetics [4]. Identification of such mutants would not have been possible with typical site-directed mutagenesis experiments.

For this thesis, I wish to tackle the global pandemic of antibiotic/antimicrobial resistance using the power of molecular dynamics simulations [13, 14]. I intend to attack this problem from two separate angles, by targeting two different proteins systems in pathogenic Gram-negative bacteria. The first target is one of the most important and oldest sources of resistance: multidrug efflux pumps [15]. These large protein complexes, which span the cellular envelope, bind antimicrobial drugs and other harmful compounds in the periplasm and cytoplasm of the cell, load them into their central "barrels," and "shoot" them out of the cell. By targeting these pumps and disabling them with so-called efflux pump inhibitors (EPIs), we can revive the efficacy of drugs for which bacteria had previously developed resistance [16, 17]. My work analyzes the flexibility of the central piece of the complex, the so-called membrane fusion protein, which binds the two membrane-bound components to complete the pump assembly complex. By manipulating its conformational landscape via new drug molecules, I aim to obstruct pump assembly and/or function.

The second target are autotransporter proteins. This proteins primarily act as virulence factors, which the bacteria use to attack other cells. The unique characteristic of autotransporters that I wish to exploit is their structure: almost all known autotransporters, across the bacterial kingdom, possess a particular helical structure that has been linked to how the virulence factors are secreted from cell [6]. Its thought that the cell uses the folding of this structure as an energy source to drive secretion as no other traditional energy sources are available at its outer membrane [5]. If we can somehow halt or stall the folding process of these autotransporters, we can prevent secretion of the virulence factors and prevent infection. By targeting a such a ubiquitous structure, the hope is that novel drugs will be effective against a broad range of bacteria and that it will be difficult for resistance to develop. My work aims to quantify the folding energy of this structure, characterize its folding pathways, and identify the essential elements of folding that could be target with new drugs.

CHAPTER 2
INTRODUCTION AND BACKGROUND

2.1 Protein Structure and Protein Folding

Proteins are the workhorses of the cell. They perform or are involved in almost every process that enables an organism to survive and reproduce. Proteins are essentially long polymer chains, called peptides, composed mostly of a collection of 20 amino acids (see Fig. 2.1), connected together via a peptide (C–N) bond. Each amino acid in the peptide chain is referred to as a "residue," and contains a common backbone and a residue-specific side chain. The chemical properties of the side chains can vary widely, from hydrophobic residues such as valine, leucine, and isoleucine, to charged residues such glutamine, arginine, and lysine. Their physical properties can also vary, such as the extremely flexible glycine or the rigged proline. Other residues enable more specific interactions, such as the base-stacking capabilities of tryptophan, phenylalanine and histidine, or the creation of disulfide bonds between neighboring cysteines. A peptide becomes a protein when all of these individual residues and interactions work in concert to perform a specific function within the cell.

A protein's structure the three-dimensional coordinates of the atoms that comprise the peptide is vital to its function within an organism [18]. Major advancements in the determination of these structures occurred in the 1950s. After initial X-ray diffraction experiments began to give insights into an ordered arrangements of the residues of proteins, in 1951 Pauling predicted two important structural motifs: helices and sheets, which he referred to as α and β structures, respectively [19, 20]. (Examples of these structures can be seen in Figure 2.2.) Although his prediction of the α-helix was off from what we now know to be the most prevalent of helices, α-helices and β-sheets have been found to be

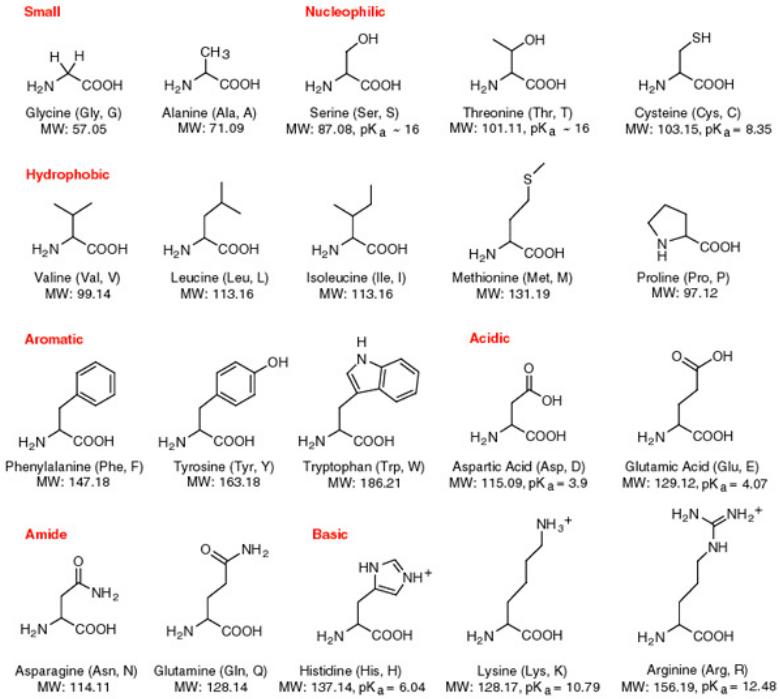
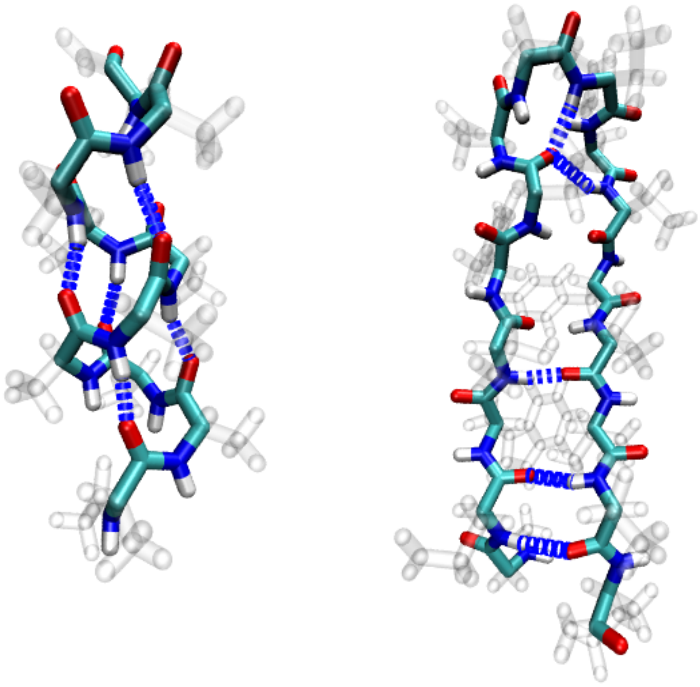


Figure 2.1: *Amino acids.* The 20 common amino acids, grouped by chemical properties.

the two most prevalent secondary structures in all proteins [21]. (Pauling would later win the Nobel Prize in 1954 for his work on the nature of the chemical bond and its relation to chemical structure). After Paulings work, the first high resolution crystal structures of hemoglobin and myoglobin were published in 1958 by Perutz [22] and Kendrew [23], respectively. This marked the first time we were able to determine and visualize a protein's entire three-dimensional structure and sparked a whole new wave of research in structural molecular biology, with Perutz and Kendrew being jointly awarded the Nobel Prize in 1962 for their work.

The idea that proteins have a distinct structure leads to an obvious question: how does it obtain that structure? This was a question posed by Levinthal at a spectroscopy conference in Illinois in 1969 [24]. He proposed the following paradox: if there are on the order of 9^{100} possible configurations (a 100-residue protein with 3 possible conformations for the ψ and ϕ dihedrals for each residue) for the typical protein, then it would take an astronomically long time for a protein to randomly sample all of these configurations until it found the



(a) (b)

Figure 2.2: *Protein secondary structures.* (a) α-helix of Ala_{10}. (b) β-hairpin of GB1. Backbone hydrogen bonds shown as dashed blue lines.

right one. He proposed that a protein may only sample a small fraction of that configuration space before it finds its proper form. However, even sampling a small fraction of that space – say 9^{10} conformations – would take too long to completely sample randomly.

Therefore, there is probably some sort of pathway along an energy surface that a protein follows towards an energy minimum where the native state resides. This is indeed how proteins fold, and much work since then has gone into determining this pathway for a variety of proteins, and if there is any generalized folding pathway that exists for all, or the majority of, proteins. The most prevalent theory is that a hydrophobic core exists at the center of most proteins, the collapse of which is the first step in the folding process. As the core collapses (or after it has fully collapsed), the internal hydrogen bonds or other contacts which maintain the structure of the protein begin to align into the native structure. Surface residues, which do not contribute to the structure, then maintain the protein interaction network within the cell. Another important driving force of folding is cooperativity.

Cooperativity occurs when the formation of a local portion of the structure induces the formation of the entire structure. This can both speed up folding and force proteins to avoid other non-native folds which may be harmful, or at the very least wasteful, to the cell. An example of cooperativity is Zimm-Bragg theory [25], which states that the formation of a single turn in an α-helix lowers the free energy barrier for the next turn to form, creating a cascading event in which each subsequent turn forms immediately after the previous one.

It is important to understand how proteins obtain their structure because protein misfolding can be very toxic to cells. Not only are misfolded proteins wasteful to the cell – in terms of time, resources, and energy – but a misfolded protein can disrupt the protein interaction network maintained by proteostasis. Additionally, misfolded proteins can also aggregate within the cell, and possibly cause other correctly folded proteins to misfold leading to further aggregation. One of the main theories for the cause of Alzheimers disease is the misfolding of the Aβ peptide, which can aggregate to form β-sheets with other misfolded Aβ, creating long fibrils [26]. These fibrils then aggregate to form plaques in the brain. One of the ways it is believed that Aβ is able to aggregate is that a misfolded Aβ can cause correctly folded Aβ to unfold, creating a nucleation event which sparks the aggregation [27]. Many prion diseases act in a similar way, with the misfolding and aggregation of prion proteins to form fibers in the brain, resulting in neurodegeneration [28].

2.2 Molecular Dynamics

Molecular dynamics (MD) is a why of simulating molecular systems by using a molecular mechanics (MM) force field, which approximates the full quantum mechanical treatment

of the molecule(s). A typical all-atom additive potential is given by

$$
\begin{aligned}
U_{\mathrm{MM}} = & \sum_{\text{bonds}} k_b(b-b_0)^2 + \sum_{\text{angles}} k_\theta(\theta-\theta_0)^2 + \sum_{\text{dihedrals}} k_\phi[1+cos(n\phi-\delta)] \\
& + \sum_{\text{impropers}} k_\omega(\omega-\omega)^2 + \sum_{\text{Uray}-\text{Bradley}} k_u(u-u_0)^2 \\
& + \sum_{\text{nonbonded}} \epsilon \left[\left(\frac{R_{\min,ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{\min,ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\epsilon r_{ij}}.
\end{aligned} \tag{2.1}
$$

For the CHARMM force field, the parameters are fitted to quantum chemistry calculations and experimental data. More specifically, partial charges are calculated using interactions with water molecules. The "bonds" term is for the bond lengths between atoms. The "angles" term is for the angles between bonds. The "dihedrals" term is for rotations about bonds. The "impropers" term governs the tetrahedral shape of 4-atom groups, typically methyl (–CH_3) groups. The "Urey-Bradley" term is a next-nearest-neighbor interaction, which is an approximation of many-body quantum effects for bonded atoms. (Although this is a poor approximation of quantum many-body effects, Urey-Bradley terms are kept for hitorical purposes.) The first "nonbonded" term is a Lennard-Jones (6-12) potential, which approximates the Pauli repulsive force due to overlapping electron orbitals at short distances (r^{-12} term) and the van der Waals attraction at large distances (r^{-6} term). The second "nonbonded" term is the Coulomb potential.

An important aspect of these types of additive force fields is that they are not polarizable, i.e. the charges are not affected by their environment. Polarizable force fields, on the other hand, have such an ability. Examples of polarizable force fields include AMOEBA [29], which introduces higher multipole terms in the Coulomb potential; the CHARMM fluctuating charge model [30, 31]; and full quantum mechanical descriptions, such as X-Pol [32]. One common thread among these polarizable models is that they are orders of magnitude slower than traditional, non-polarizable models. The CHARMM Drude polarizable force field is of particular interest due to its relatively low computational cost

compared to other polarizable force fields [33]. Based on the classical drude oscillator, each heavy (i.e. non-hydrogen) atom is split into a positive "parent" atom, representing the positively charged nucleus, and a negative "drude" particle, representing the negatively charged electron cloud surrounding the nucleus (Fig. 2.3). The drude particle is attached to the parent atom via a stiff spring with equilibrium value set to the center of the parent atom. The drude particles are given a mass of 0.4 AU and are kept at a temperature close to absolute zero (1 K for my simulations) for stability purposes. Drude particles do not interact via the Lennard-Jones potential, so an infinite hard wall potential is usually added between 0.2 to 0.25 Å to avoid the so-called "polarization catastrophe," wherein the drude particle of one atom falls into the Coulomb potential of a neighboring atom. Converting from the non-polarizable system to a drude polarizable system effectively doubles the total number of atoms in the system, and simulating the drude model requires the timestep be cut in half. So computationally, the efficiency of the drude model is about $1/4$ that of the non-polarizable CHARMM force fields, a great improvement over other polarizable force fields.

The force on the i-th atom is then generated by the gradient of the MM potential, i.e. $\boldsymbol{F}^{(i)} = -\nabla^{(i)} U_{\mathrm{MM}}$, where $\nabla^{(i)}$ is the gradient with respect to $\boldsymbol{x}^{(i)}$. In order to couple the system to a heat bath, a Langevin thermostat is used, resulting in a set of stochastic differential equations given by

$$m_i \ddot{\boldsymbol{x}}^{(i)} = -\nabla^{(i)} U_{\mathrm{MM}} - \gamma m_i \dot{\boldsymbol{x}}^{(i)} + \sqrt{2\gamma k T m_i}\boldsymbol{R}^{(i)}(t), \tag{2.2}$$

with γ set to $5\ ps^{-1}$ for simulations described here, T the target temperature, k the Boltzmann constant, and $\boldsymbol{R}$ a completely uncorrelated random variable with mean 0:

$$\begin{aligned} \left\langle \boldsymbol{R}^{(i)}(t) \right\rangle &= \mathbf{0} \\ \left\langle \left(\boldsymbol{R}^{(i)}\right)_j (t) \left(\boldsymbol{R}^{(i')}\right)_{j'} (t') \right\rangle &= \delta_{ii'}\, \delta_{jj'}\, \delta(t - t'). \end{aligned} \tag{2.3}$$

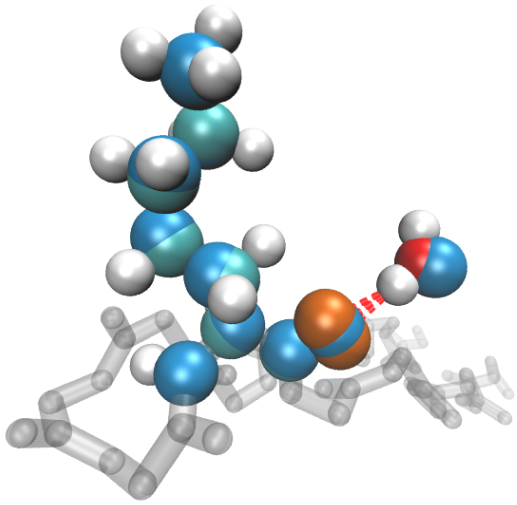


Figure 2.3: *CHARMM drude polarizable model.* Drude polarizable model of lysine (left) and the SWM4-NDP water model (right), in the van der Waals representation. Atom colors are as follows: (cyan) carbon, (red) oxygen, (white) hydrogen, (dark blue) nitrogen, (light blue) drude particles, (orange) lone pair electrons. A hydrogen bond between the carbonyl oxygen of the lysine backbone and the oxygen of the water molecule is shown as a dashed red line.

Molecular dynamics has been used to simulate systems ranging from small, isolated peptides and proteins [34] to large membrane-bound systems involving many components (including lipids, proteins and nucleic acids) [35] to entire virus capsules [36]. Not only can we simulate a wide range of systems sizes, but due to the highly parallizable nature of molecular dynamics (where calculations for small groups of atoms can be performed independently), with dedicated supercomputers such as Anton [37], we can now simulate these systems into the millisecond time range [38]. This has enabled us to view and measure events that occur closer to biological timescales, such as the folding, unfolding and refolding of entire proteins, not just small structural motifs [39].

2.3 Free Energy Landscapes from Molecular Dynamics Simulations

The benefit of using molecular dynamics to study biological systems is that there is much greater control of the system as well as a greater scope of analysis. The dynamics and energetics of every single atom are at your disposal. This is not only useful for monitoring and measuring your system, but you can also directly modify it in any way you choose. Free energy methods for MD capitalize on these capabilities to enhance the efficiency of free energy calculations. Without any external bias on the system, achieving an accurate free energy landscape can difficult due to the presence of high energy barriers, which make the probability of crossing these barriers very low on simulation timescales. Even if a barrier crossing is observed during a simulation, sampling of the transition state at the peak of that barrier will be very poor as the system spends very little time there, leading to a very slow rate of convergence of the free energy.

One way to overcome these inefficiencies is to apply external biases to the system to force it to cross barriers and spend more time sampling transition states. This is referred to as "enhanced sampling." The free energy is then calculated as a potential of mean force (PMF), where the external biases are then removed during the analysis stage. A simple method is to apply a force to push the system along the reaction pathway of your choice.

For protein folding, this can be achieved by applying a constant (ideally slow) velocity force along a particular reaction coordinate, ξ, using steered molecular dynamics (SMD). The PMF can then be calculated using Jarzynski's equality:

$$e^{-F/k_BT} = \left\langle e^{-W/k_BT} \right\rangle, \tag{2.4}$$

where the Boltzmann weight of the equilibrium free energy, G, is simply the ensemble average of the Boltzmann weight of the non-equilibrium work, W [40].

There are another class of of free energy methods that aim to flatten the energy landscape so that the system simply diffuses along the reaction coordinate. One such method is metadynamics [41], which consists of adding Gaussian hills to the region of the free energy surface that has already been visited, pushing the system out of potential wells towards transition states. Eventually enough Gaussians are added such that the free energy profile is effectively flat, allowing the system to freely diffuse along the reaction coordinate. Adaptive biasing forces (ABF) works in similar manner [42, 43, 44]. Instead of adding Gaussian hills to previously traversed regions, ABF simply adds cancels out the forces in that region. First, the reaction coordinate is binned into small $\Delta\xi$'s. In each bin the average force is calculated, and as the number of samples in a bin reaches some threshold value (called "fullSamples" in NAMD), the biasing force is slowly ramped up to the negative of the average force in that bin. After long enough sample, ABF will also produce an effectively flat free energy surface for the system to diffuse across.

The method I use in the majority of my simulations is umbrella sampling (US). US allows you to sample different regions of the reaction coordinate space independently, allowing for high parallization of the simulations, which increases efficiency. Similarly to the ABF method, you partition your reaction coordinate space into smaller windows. For each window, a state corresponding to that window is created (often by SMD or ABF). The state is then restrained to only sample a small region around the center of each window, allowing

for some overlap with the neighboring windows. This is typically achieved by applying a harmonic restraining force to the reaction coordinate, with equilibrium value given by the center of the window. The Weighted Histogram Analysis Method (WHAM) [45] is then used to reconstruct the PMF by examining the histograms from each window, removing the bias from the harmonic restraint, and evaluating the overlapping regions between windows.

Another potential problem that can arise in free energy calculation methods is that of orthogonal degrees of freedom. Free energy methods often rely on you to choose the reaction coordinate, so it must be a priori known. True reaction pathways, however, tend to be very complex and multidimensional [46, 47, 48, 49], so your enhanced sampling technique may not be able to sample the true transition state and find new free energy minima if you've only chosen one or two reaction coordinates along which to sample. Replica exchange molecular dynamics (REMD) allows the system to traverse orthogonal degrees of freedom by allowing neighboring windows to swap their harmonic restraints. Originally developed as a temperature exchange, or parallel tempering, method [50], it has been adapted to allow for the exchange of potential energy parameters (often referred to as Hamiltonian exchange) [51]. Replica exchange umbrella sampling (REUS) is more efficient than traditional US as orthogonal degrees of freedom are allowed to relax during window exchanges.

CHAPTER 3

STATISTICAL MECHANICS AND ENHANCED SAMPLING TECHNIQUES

3.1 Free energy and probabilities

For systems at constant pressure and temperature, the NPT ensemble, the Gibbs free energy of a system is given by

$$G = H - TS \tag{3.1a}$$

$$= U + PV - TS \tag{3.1b}$$

where H is the enthalpy, S is the entropy, T is the temperature, U is the potential energy, P is the pressure, and V is the volume. The probability of state i, p_i, is given by its Boltzmann weight:

$$\rho_i = e^{-G_i/kT}, \tag{3.2a}$$

$$p_i = \frac{\rho_i}{\sum_j \rho_j} = \frac{\rho_i}{Z}, \tag{3.2b}$$

where k is the Boltzmann constant, with a value of 0.001987 kcal/mol. Z, the sum of all the Boltzmann weights, is called the partition function. From Eq. 3.2, one can calculate the free energy of each state by measuring its probability. This can be done by simply counting the number of states that fall into state i over the course of the trajectory, $\{r(t_1), r(t_2), r(t_3), \ldots\}$, and dividing by the total length of the trajectory. Due to the limited timescales in a typical MD simulation, it is often critical to introduce additional forces into the system to drive it to a particular state. If we introduce biasing potentials, ΔU to the system ($U' = U + \Delta U$), then the free energy of the unbiased system is simply the biased free energy minus the bias potential. Alternatively, we can calculate the unbiased

probabilities by reweighting the counts according to Eq. 3.2a:

$$\text{count}(t_i) = e^{\Delta U[r(t_i)]/kT}. \tag{3.3}$$

3.2 Adaptive biasing forces (ABF)

Adaptive biasing forces (ABF) is a dynamic bias force that attempts to smooth out the free energy surface (FES) over time [42, 43, 44, 52]. First the reaction coordinate, ξ, which is generally a function of many atomic coordinates, x_k, is broken down into small bins. Over time as the system enters and re-enters a bin, ξ^*, the average force on the reaction within that bin is calculated. To see how the average force is related to the free energy, first we start with the definition of the free energy in integral form:

$$G(\xi^*) = -\frac{1}{kT}\int e^{-H(\boldsymbol{p},\boldsymbol{x})/kT}\delta(\xi(\boldsymbol{x}) - \xi^*)d\boldsymbol{p}d\boldsymbol{x}. \tag{3.4}$$

The next step is referred to as thermodynamic integration (TI), where we measure the derivative of G instead of G directly:

$$\frac{dG}{d\xi}(\xi^*) = \frac{\int \frac{\partial H}{\partial \xi} e^{-H/kT}\delta(\xi(\boldsymbol{x}) - \xi^*)d\boldsymbol{p}d\boldsymbol{x}}{\int e^{-H/kT}\delta(\xi(\boldsymbol{x}) - \xi^*)d\boldsymbol{p}d\boldsymbol{x}} = \left\langle \frac{\partial H}{\partial \xi}\middle| \xi^* \right\rangle. \tag{3.5}$$

By noting that the derivative of the enthalphy, H, is just the force, we arrive at the average force to be used as our bias:

$$\frac{dG}{d\xi}(\xi^*) = -\left\langle F_\xi | \xi^* \right\rangle. \tag{3.6}$$

Eq. 3.5 can written in terms of the potential energy, U, as follows [52]:

$$\frac{\partial G(\xi^*)}{\partial \xi} = \left\langle \frac{\partial U}{\partial \xi} + \frac{\partial \log|J|}{\partial \xi}\middle| \xi^* \right\rangle, \tag{3.7}$$

where J is the Jacobian matrix of ξ with respect to the atomic coordinates. Eq. 3.7 can be cumbersome to calculate in practice due to the spatial derivatives involved. A faster,

time-derivative-based formula was later derived that can be much more easily calculated on-the-fly during an MD simulation [52]:

$$\frac{\partial G(\xi^*)}{\partial \xi} = -\left\langle \frac{d}{dt}\left(m_\xi \frac{d\xi}{dt}\right) \middle| \xi^* \right\rangle, \tag{3.8}$$

where $m_\xi^{-1} = \sum_k m_k^{-1} \partial^2 \xi / \partial x_k^2$, and m_k is the mass of the kth particle. Now, Eq. 3.8 is akin to Newton's Second Law, which MD programs already utilize to calculate the molecular mechanical forces. For multiple reaction coordinates, $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3, \ldots)$, the ABF biasing force is calculated as follows:

$$\boldsymbol{F}_k(N_{\text{step}}, k) = \frac{1}{\mathcal{N}(N_{\text{step}}, k)} \sum_{i=1}^{\mathcal{N}(N_{\text{step}},k)} \boldsymbol{F}_i(t_i^k), \tag{3.9a}$$

$$\boldsymbol{F}_i(t_i^k) = \frac{d}{dt}\left(M_{\boldsymbol{\xi}} \frac{d\boldsymbol{\xi}}{dt}\right)\Big|_{t_i^k}, \tag{3.9b}$$

where $\boldsymbol{F}_i(t_i^k)$ is the ith force sample when $\boldsymbol{\xi}$ is in bin k and t_i^k is the time at which sample i was collected.

The free energy, $G(\boldsymbol{\xi})$, is then recovered by numerically integrating Eq. 3.6 over all the bins of $\boldsymbol{\xi}$. One usually sets the number of samples within each bin that must accumulated before a good average can be calculated. The force is slowly turned on from 0 to the full average over the course of this preset sampling threshold to maintain quasi-adiabaticity and minimize the amount of non-equilibrium work added to the system.

3.3 Metadynamics

Another time-dependent biasing method is metadynamics [41]. Instead of calculating an average force over time, metadynamics adds a history-dependent potential to the Hamiltonian:

$$V_{\text{meta}}(\xi(t)) = \sum_{t'=\delta t, 2\delta t, \ldots}^{t'<t} W \prod_i^{N_{\text{cv}}} e^{-\frac{(\xi_i(t)-\xi_i(t'))^2}{2\sigma_{\xi_i}^2}}, \tag{3.10}$$

where W, σ_{ξ_i}, and δt are parameters chosen to balance the speed of exploration of the reaction coordinate coordinate space with the accuracy of the calculated FES, which is approximated by:

$$G(\xi) \simeq -V_{\mathrm{meta}}(\xi) + K, \tag{3.11}$$

where K is an arbitrary constant. Essentially, the system keeps adding small Gaussian hills to the potential energy as the system moves about the reaction coordinate space to push the system out of areas it has already explored. Once a particularly energy well as been sufficiently explored, eventually the well will fill up with Gaussians to the point that the system eventually diffuses to the next well, where the process begins again. After all the wells have been filled, the resulting potential energy surface will be roughly flat, and the system can easily diffuse across the entire FES.

Because the Gaussian hills in Eq. 3.10 get added regardless of how often an area is explore, the average free energy calculated with Eq. 3.11 will never truly converge to the true free energy, but instead will oscillate around the true value as the system travels back and forth across the FES. To solve this problem, the weights of the Gaussian hills can decrease if a particular area has already be sufficiently sampled [53]. The weight, W, in Eq. 3.10 is now dependent on $\boldsymbol{xi}$ and t and takes the following form:

$$W \rightarrow \omega e^{-V_{\mathrm{meta}}(\xi,t)/k\Delta T}\delta t, \tag{3.12}$$

where $V_{\mathrm{meta}}(\xi, t)$ is the value of the bias potential at ξ up to time t, ΔT is a temperature that describes how quickly the bias force decays as more samples accumulate, and $\omega \cdot \delta t$ is some chosen initial height of the Gaussians. With these "well-tempered" weights, Eq. 3.11 converges smoothly to the true free energy [53].

3.4 Umbrella sampling (US)

The method of "umbrella sampling" (US) was originally proposed in the context of Monte Carlo simulations for fluid dynamics by Torrie and Valleau in 1977 [54]. The idea was to calculate energy distributions or ensemble averages from a "system of interest" by instead examining a reference system, assuming the relationship between the two systems was analytically known. In molecular dynamics, it now refers to Eq. 3.3 where the bias, ΔU, is a harmonic potential. More specifically, US involves many simulations with different harmonic biasing potentials that are all centered at different points along your reaction coordinate. By combining the trajectories from all simulations, one can cover the entire reaction coordinate space of interest using many short simulations instead of one long simulation. This high degree of parallelization is one of US's many advantages over other enhanced sampling techniques such as ABF or metadynamics. It also ensures a more uniform coverage of the entire reaction coordinate space. ABF and metadynamics are useful for first exploring the space, however, so they can used to seed the windows of an US simulation. The data from each system is later combined together to recreate the FES. This is typically done by the weighted histogram analysis method (WHAM), which will be discussed in the next section, although other methods are available, such as the dynamical weighted histogram analysis method (DHAM) [55], the transition-based reweighting analysis method (TRAM) [56], umbrella integration (UI) [57], and linear regression [58].

3.4.1 Weighted histogram analysis method (WHAM)

The weighted histogram analysis method (WHAM) recreates an unbiased FES from a collection of K biased simulations, each with some combination of L biasing potentials, U_i, and temperature T_i, by reweighting the trajectory counts according to Eq. 3.3. The Hamil-

tonian and free energy of each simulation k can be written as

$$H_{\{\lambda\}}(\boldsymbol{x}) = H_0(\boldsymbol{x}) + \sum_{i=1}^{L} \lambda_i U_i(\boldsymbol{x}), \tag{3.13a}$$

$$G_{\{0\},\beta} = G_{\{0\},\beta} - \sum_{i=1}^{L} \lambda_i U_i(\xi) + C(\{\lambda\}, \beta), \tag{3.13b}$$

where $\{\lambda\}$ is a set of weights within each simulation for each of the biasing potentials, C is an arbitrary constant, and $\beta = 1/kT$. For a typical one-dimensional umbrella sampling simulation, $K = L$, $U_i(\xi) = \frac{1}{2}k_i(\xi - \xi_{0,i})^2$ for set of ξ_0's, and $\lambda_i = \delta_{i,j}$ for simulation j.

Eq. 3.13 generates K arbitrary constants. These constants must be chosen such that the FES from each simulation can be fitted together to create a single smooth surface. By minimizing the error between the probability distributions, one obtains the so-called "WHAM" equations [45]:

$$P_{\{\lambda\},\beta}(\{U\}, \xi) = \frac{\sum_{k=1}^{K} N_k(\{U\}, \xi) \exp\left(-\beta \sum_{j=1}^{L} \lambda_{j,k} U_j\right)}{\sum_{m=1}^{K} n_m \exp\left(f_m - \beta_m \sum_{j=1}^{L} \lambda_{j,m} U_j\right)}, \tag{3.14a}$$

$$\exp(-f_j) = \sum_{\{U\},\xi} P_{\{\lambda\}_j,\beta_j}(\{U\}, \xi), \tag{3.14b}$$

where $N_i(\{U\}, \xi)$ is value taken by the histogram at values $\{U\}$ and ξ during simulation i, $\{U\}$ is the set of biasing potentials, n_i is the total number of points taken from simulation i, and f_i is the total (dimensionless) free energy of the ith simulation ($f_i = \beta_i G_i$). To improve the accuracy, Eq. 3.14b is usually calculated directly from the simulation data by

$$\exp(-f_i) = \sum_{k=1}^{K} \sum_{t=1}^{n_k} \frac{\exp\left[-\beta_i \sum_{j=1}^{L} \lambda_{j,i} V_{j,t}^{(k)}\right]}{\sum_{m=1}^{K} n_m \exp\left[f_m - \beta_m \sum_{j=1}^{L} \lambda_{j,m} V_{j,t}^{(k)}\right]}, \tag{3.15}$$

where $V_{i,t}^{(l)}$ is the value of potential V_i at snapshot t in the l simulation. Eq. 3.15 is then solved iteratively, starting with an initial guess for the f_i's, typically setting them all to zero.

For a large number (>1000) of simulations, also called windows, this iterative procedure can take a prohibitive amount of time. Not only are there more f_i's that need to be solved, the number of terms on the right hand side of Eq. 3.15 also increases with the number of windows. To increase the convergence rate of this iterative scheme, one can implement direct inversion in the iterative subspace (DIIS) [59].

In an iterative scheme, one aims to satisfy the following target equation:

$$R_i(\boldsymbol{f}) = 0 \quad \forall i = 1, \ldots, K. \tag{3.16}$$

R_i are residual functions, equal to $-\log \mathcal{Z}_i(\boldsymbol{f}) - f_i$, where $\boldsymbol{f}$ is a vector containing all the f_i's, and $\mathcal{Z}_i(\boldsymbol{f})$ is given by the r.h.s. of Eq. 3.15. The error on $\boldsymbol{f}$ is given by $||\boldsymbol{R}(\boldsymbol{f})||$, and $\boldsymbol{R}(\boldsymbol{f})$ should optimally point in the direction that minimizes the error on $\boldsymbol{f}$. In a direct iterative scheme, this is done by updating the vector $\boldsymbol{f}$ with $\boldsymbol{f} - \boldsymbol{R}(\boldsymbol{f})$, and recalculating $\boldsymbol{R}(\boldsymbol{f})$. This procedure is repeated until $||\boldsymbol{R}(\boldsymbol{f})|| < \delta$, where δ is some numerical tolerance. $\boldsymbol{R}(\boldsymbol{f})$, however, does not always point in the proper direction, or does not have the right magnitude, to reduce the error of $\boldsymbol{f}$. Since $||\boldsymbol{R}(\boldsymbol{f})||$ is still a reliable measure of the error of $\boldsymbol{f}$, we can look for a different $\hat{\boldsymbol{f}}$ which does minimize the error.

In the DIIS scheme, the vector $\hat{\boldsymbol{f}}$ is a linear combination of basis vectors, $f_1, f_2, \ldots, f_M$, where M can be much less than the number of simulations, K. We now have M residual vectors, $\boldsymbol{R}_i = \boldsymbol{R}(\boldsymbol{f_i})$, with the full residual vector a linear combination of the individual residual vectors, $\hat{\boldsymbol{R}} = \sum_{i=1}^{M} c_i \boldsymbol{R}_i$. We wish to minimize the error, $||\hat{\boldsymbol{R}}||$, with the restraint

$$\sum_{i=1}^{M} c_i = 1. \tag{3.17}$$

This can be solved using a lagrange multiplier, λ, resulting in the M equations,

$$\sum_{i=1}^{M}(\boldsymbol{R_i}\cdot\boldsymbol{R_j})c_i = \lambda, \tag{3.18}$$

that need to be solved simultaneously with Eq. 3.17. λ is another parameter, along with the c_i's, that needs to be determined. Since Eq. 3.16 should be nearly linear near the true solution, $\boldsymbol{f}^*$, $\hat{\boldsymbol{f}} = \sum_{i=1}^{M} c_i \boldsymbol{f}_i$ should be close to true solution. After $\hat{\boldsymbol{f}}$ is determined, a new basis vector is found by

$$\boldsymbol{f}^{(n)} = \hat{\boldsymbol{f}} + \alpha\hat{\boldsymbol{R}}(\hat{\boldsymbol{f}}). \tag{3.19}$$

To determine the basis set of vectors, $\boldsymbol{f}_i$, first start with a single vector (typically the initial guess of zero for all f's). Usually we would like to limit our basis to some size, M. So as new basis vectors are formed, old vectors may be discarded. The scheme for building up the basis in this thesis is the one used in Zhang et al [59]. First determine the basis with maximal error, $\boldsymbol{f}_{\max}$. If $\boldsymbol{f}^{(n)}$ produces an error less than $\boldsymbol{f}_{\max}$, $\boldsymbol{f}^{(n)}$ is added to the basis, or if the basis is full, $\boldsymbol{f}_{\max}$ is replaced by $\boldsymbol{f}^{(n)}$. Otherwise, $\boldsymbol{f}_{\max}$ is removed from the basis. If the basis then becomes empty, the basis is restarted with $\boldsymbol{f}^{(n)}$.

3.4.2 Replica exchange molecular dynamics (REMD) and umbrella sampling (REUS)

A common problem for most enhanced sampling simulations where a reaction coordinate is being biased is the relaxation along orthogonal degrees of freedom (ODF), especially if their relaxation times are slower than for your chosen reaction coordinate. One popular method for speeding up the relaxation of ODEs is is the use of high temperatures. Replica-exchange molecular dynamics (REMD) takes advantage of high temperatures while also giving useful information at low temperatures. It works similarly to umbrella sampling in that you have multiple replicas of the system at different temperatures (temperature windows instead of reaction coordinate windows as in US) [50]. The key is to allow an exchange of information between each window. This is accomplished by periodically ex-

changing conformations of the system between neighboring windows. Just as in a Monte Carlo walk through temperature space, the Metropolis criterion is used decide when window i and j should exchange their conformations:

$$P(i \leftrightarrow j) = \min\{1, \exp(-\Delta E)\}, \tag{3.20a}$$

$$\Delta E = (\beta_i - \beta_j)(E(\boldsymbol{q}_i) - E(\boldsymbol{q}_j)), \tag{3.20b}$$

where replicas i and j are at temperatures T_i and T_j and have conformations $\boldsymbol{q}_i$ and $\boldsymbol{q}_i$, respectively, and $P(i \leftrightarrow j)$ is the probability of exchanging conformations between replicas i and j. If we let $T_i > T_j$ ($\beta_i < \beta_j$), then the Metropolis criterion examines the energy difference between conformations i and j. If $E(\boldsymbol{q}_i) > E(\boldsymbol{q}_j)$ ($\Delta E < 0$), i.e., the higher temperature replica is in a lower energy conformation, then the system always switches them. Otherwise, the switch occurs with some probability given by the Boltzmann weight of the energy difference. Since higher temperature replicas can more easily cross energy barriers, they can more quickly find lower energy states that may be kinetically inaccessible at lower temperatures. The replica exchange scheme allows this those lower energy conformations to "diffuse" down temperature space to the target temperature. A common swapping scheme, known as the "even-odd" scheme, is to allow only neighboring temperatures to swap. Occasionally, you can also allow all replicas to randomly swap.

Replica exchange umbrella sampling (REUS) works in a very similar manner, except the Metropolis criterion is used to swap conformations between neighboring US biasing potentials instead of temperatures [51]. This gives rise to a modified equation for ΔE in the Metropolis criterion:

$$\Delta E = \beta[(E_i(\boldsymbol{q}_i) + E_j(\boldsymbol{q}_j)) - (E_i(\boldsymbol{q}_j) + E_j(\boldsymbol{q}_i))], \tag{3.21}$$

where E_i and E_j are the energies from ensembles with biasing potentials $U_i(\xi)$ and $U_j(\xi)$, respectively. For REUS, $\Delta E < 0$ when the combined energies of replicas i and j are lower

after swapping their conformations. Whereas the goal of REMD is to get the lowest energy conformations in the target temperature, the goal of REUS is to lower the combined energy of all the simulations since US covers the entire reaction coordinate space all at once. The REUS method takes advantage of movement along ODFs from different areas of the reaction coordinate space with low energy barriers along those orthogonal coordinates to diffuse to regions with higher barriers. For optimal sampling and faster convergence of the PMF, the exchange probability (Eq. 3.20a) should be $\sim$0.4 between all neighboring windows [60].

3.4.3 Self-learning adaptive umbrella sampling (SLUS)

The umbrella sampling method can also be used to explore the FES in much the same way ABF and metadynamics can. Bernèche and coworkers developed the Self-learning adaptive umbrella sampline (SLUS) method to search only along low-energy paths [8]. A free-energy cutoff must be chosen *a priori* such that only regions below this cutoff will be explored. Some biological and/or chemical intuition is therefore needed to decide the roughly what energies are physically relevant for your system.

The algorithm starts by first calculating the FES through normal umbrella sampling simulations starting from some initial set of windows. Then a new set of windows is spawned from those original windows which lie below the free energy cutoff. For this reason, a regular grid of umbrella windows is typically needed. New umbrella sampling simulations are then carried on the new windows. Starting states for these new windows can be generated by searching through the previous neighboring windows for the state which is closest to the umbrella window center. You can also search exclusively through the lowest energy neighboring window [8]. The algorithm is repeated until no new windows can be generated. You can also use a range of cutoffs where, once the algorithm stops with one cutoff, you can advance the algorithm to a higher cutoff and repeat the procedure.

CHAPTER 4

THERMODYNAMICS OF DECA-ALANINE FOLDING IN WATER

4.1 Introduction

Folding of proteins into organized three-dimensional structures capable of fulfilling a biological function is determined by the sequence of amino acids [18] and is believed to proceed hierarchically [61, 62]. The emergence of secondary-structure elements constitutes an early event in the chronology of folding [63], which prefaces the ultimate collapse into well-defined, compact, functional entities. Formation of stretches of secondary structure, the elementary bricks of the protein scaffold, therefore, represents an important milestone on the folding pathway, and a convenient framework to investigate the basic physical principles that underlie protein folding — notably how do elements of secondary structure nucleate and further propagate into an ordered structure, and to what extent is the organization of the peptide chain collective [64]. Understanding this key biological process at the theoretical level has greatly benefited from the recent development of novel, dedicated computer architectures [65] and the unbridled race to model larger proteins over longer timescales [66, 67]. Brute-force simulations have now reached a cruising speed that can fold proteins as large as one hundred amino-acid residues over tens to hundreds of microseconds [39], still at the price of substantial computational effort. A number of unbiased, all-atom simulations in explicit solvent have proven successful to illuminate the hierarchical nature of folding, shedding light on the possible pathways that connect a random coil to a functional three-dimensional structure [68, 69, 70, 71].

Substantially shorter importance-sampling [72, 73] simulations relying on simpler models consisting of short, organized peptide segments can, however, provide valuable insight into the physical and evolutionary principles that govern the intricate conformational tran-

sition of a disordered protein chain into a properly folded one [64, 74, 75, 76, 77, 78]. Among suitable candidates of secondary-structure elements for biased simulations are α–helices, the most prevalent motif observed in proteins [21], stabilized by intramolecular interactions, notably through the formation of hydrogen bond between the carbonyl moiety of the i–th residue and the amino moiety of the i+4–th residue. Owing to its noteworthy propensity to form α–helices [79], alanine has been the amino acid of predilection in theoretical investigations of conformational equilibria in short peptide segments. Alanine-rich peptides flanked by titratable residues have also been utilized abundantly at the experimental level [80] to decipher the transition pathway from a disorganized chain to a nascent chain to an ultimately folded α–helix. In particular, they were at the center of a controversy on the existence of 3_{10}–helices [81], a secondary-structure motif arising from the formation of hydrogen bonds between the i–th and the i+3–th residues of the peptide chain, conjectured to act as an observable intermediate in the conformational transition towards the α-helical state.

Turning to importance-sampling simulations naturally raises the question of an appropriate choice of a transition coordinate, capable of describing folding of the peptide chain into a well-ordered secondary structure. Even for appreciably short segments, this choice remains an intricate problem, deeply rooted in the large number of degrees of freedom that vary concurrently as the peptide chain evolves towards its native, organized conformation [82]. Much of this intricacy lies in the multidimensionality of the true reaction coordinate [46, 47, 48, 49], thwarting naive attempts to resort to a limited number of geometric variables, often of low collectivity [83]. Fruitful application of collective-variable-based methods rests in large measure upon the fragile hypothesis of timescale separation of slow degrees of freedom, in connection with the reaction coordinate, as well as all other hard, fast degrees of freedom. Mapping the free-energy landscape that underlies the folding of a short peptide, therefore, ultimately reduces to either select a few relevant collective variables, or throw into the model a plethora of order parameters to describe the multi-

dimensional transition space [84]. The daunting nature of this task explains why biased simulations of complex, intertwined conformational changes remain scarce [85].

In the present contribution, we revisit the paradigmatic capped decamer of alanine, henceforth referred to as deca-alanine. Deca-alanine has served on various occasions as a methodological proof of concept, in particular in non-equilibrium work simulations in conjunction with the Jarzynski identity [86], and equilibrium free-energy calculations relying upon the application of a time-dependent bias [42, 87]. Notwithstanding their rudimentary character, model peptides like deca-alanine offer valuable thermodynamic and kinetic information on folding, under the assumption that a reasonable, non-ambiguous transition coordinate can be designed — which is necessarily subservient to the length of the peptide chain. They also help shed light on common shortcomings of importance-sampling simulations of low-dimensionality, notably hidden barriers in orthogonal space, and have proven useful for devising remedies [87, 88, 89, 90]. Beyond their undeniable utility in methodological developments, they are also sufficiently simple to serve as models of the nascent chain in more realistic biological applications, like the coupled folding–translocation occurring in the SecY complex [35].

Here, we extend the exploration of reversible extension of deca-alanine in vacuo [42, 87] by examining how the aqueous environment reshapes the free-energy landscape that underlies folding. We find that the range of conformational states explored in water is much greater than in vacuum, making end-to-end distance a highly degenerate transition coordinate. However, by adding a second coordinate describing the helicity of deca-alanine, we demonstrate that its folding pathway in water is more intricate than in vacuum.

4.2 Methods

Simulations of deca-alanine (Ala_{10}) were performed using the 104-atom compact helical model used by Park et al. [86], capped with an acetylated N-terminus and amidated C-terminus, as a starting state, with all hydrogens defined explicitly. For simulations in

explicit water, the visualization and analysis program VMD [91] was used to place deca-alanine in a cube of 10,850 TIP3P [92] water molecules with dimensions 70 Å $\times$ 70 Å $\times$ 70 Å for a total of 32,659 atoms. Molecular dynamics simulations were carried out using NAMD 2.9 [**Phillips2005**] with the CHARMM all-atom force fields (CHARMM22/CMAP [93] and CHARMM36 [94, 95]). The temperature was fixed at 300 K using Langevin dynamics; the pressure was kept constant at 1 atm using the Langevin piston method [96]. The equations of motion were integrated using the RESPA multiple time-step algorithm with a time step of 2 fs used for all bonded interactions, 2 fs for short-range non-bonded interactions, and 4 fs for long-range electrostatic interactions. Long-range electrostatic interactions were calculated using the particle-mesh Ewald method [97]. Bonds involving hydrogen atoms were constrained to their equilibrium length, employing the Rattle algorithm [98].

PMFs were calculated using both adaptive biasing forces (ABF) [**Phillips2005**, 42, 44] and umbrella sampling (US) with the weighted histogram analysis method (WHAM) [45], utilizing the *collective variables* (colvars) module of NAMD 2.9 [87]. Two reaction coordinates were defined: (ξ) the distance from the carbonyl carbon of the backbone of the first residue to the carbonyl carbon of the last residue, and (α) the α-helical content of all 10 alanine residues as defined in the colvars module of NAMD. The α colvar is calculated using a scoring function for the backbone $i, i+4$ hydrogen bonding and the dihedral angles compared to that of a pure α-helix, normalized between 0 and 1. The default parameters for the α colvar as defined in the colvars module were used in all simulations. For 2-dimensional PMFs in water, replica-exchange molecular dynamics [51] was utilized with US (REMD-US) to increase the sampling efficiency of the entire conformational space. Integration of the 2D PMF to obtain a 1D PMF was calculated according to the following equation [99]:

$$e^{-\beta w(x)} = e^{-\beta W(x_c,y_c)} \frac{\int dy\ e^{-\beta W(x,y)}}{\int dy\ e^{-\beta W(x_c,y)}}, \tag{4.1}$$

where $\beta = (k_B T)^{-1}$, k_B is the Boltzmann constant, T is the temperature, $W(x, y)$ is the 2D

PMF, $w(x)$ the corresponding 1D PMF, and (x_c, y_c) is an arbitrary point in the collective-variable space.

Along the end-to-end distance coordinate, 20 US windows centered at $\xi = 12.5$ Å, 13.5 Å, ..., 31.5 Å were used with a force constant of 5.0 kcal/Å^2·mol for each window. Along the α-helical content coordinate, 9 US windows centered at $\alpha = 0.1, 0.2, \ldots, 0.9$ were used in vacuum and 17 US windows centered at $\alpha = 0.1, 0.15, \ldots, 0.9$ were used in water with a force constant of 500.0 kcal/α^2·mol for each window. In vacuum, US windows were simulated for 5-10 ns per window for 1-dimensional PMFs and 15 ns per window for 2-dimensional PMFs. In water, US windows were simulated for 5 ns per window for 1-dimensional PMFs and 20 ns per window for 2-dimensional PMFs. The first 1-2 ns were not included in the PMF calculations to ensure the system was in equilibrium. ABF simulations were run for 50-100 ns in total. All ABF simulations used a threshold of 500 samples ("full samples") prior to the application of the bias, unless noted otherwise. Starting states along each reaction coordinate were generated using either steered molecular dynamics (SMD) or from 1-dimensional unrestrained ABF trajectories.

4.3 Results

4.3.1 One-dimensional PMFs

To examine the efficacy of our methods, we first determined the PMF of deca-alanine in vacuum using the end-to-end distance reaction coordinate, denoted ξ. Using both the US and ABF approaches (see Methods), we calculated the PMF with the CHARMM22/CMAP and CHARMM36 force fields. The two approaches yield nearly identical free-energy profiles for both force fields (Fig. 4.1). Examination of the simulation trajectories shows that both methods produced only the accordion-like folding/refolding mechanism as shown in Fig. 4.2, where unfolding begins at one end of the peptide and propagates to the other end, suggesting a cooperative folding mechanism [25].

The results of the CHARMM36 force field agree quite well with previously reported

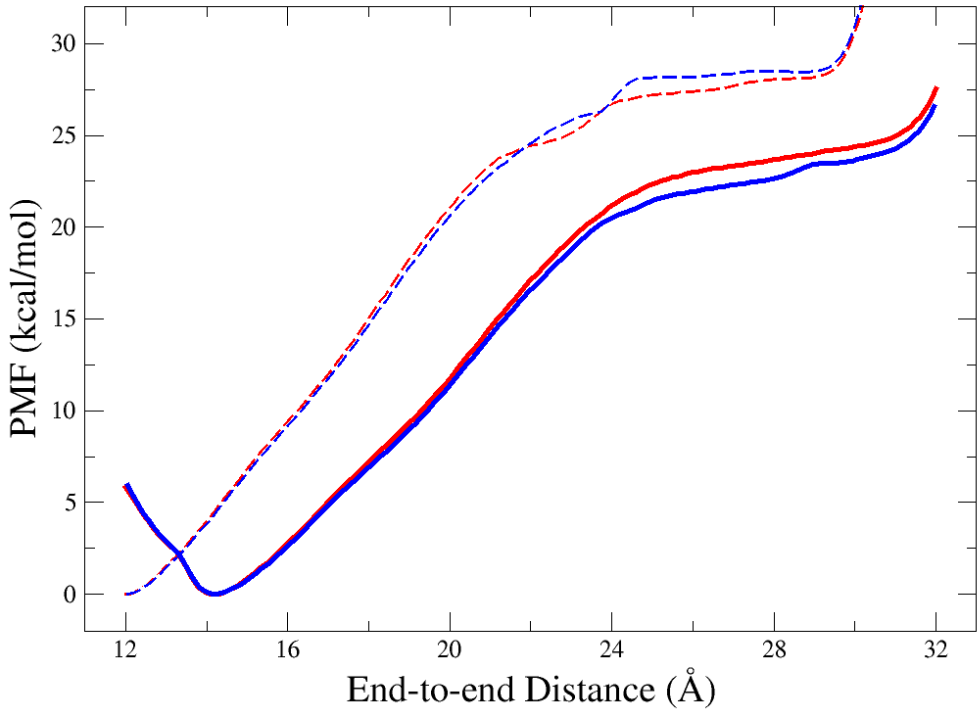


Figure 4.1: *One-dimensional PMF of Ala_{10} in vacuum.* Calculated using the distance from the N-terminus carbonyl carbon to the C-terminus carbonyl carbon as the reaction coordinate. Red lines represent calculations using ABF and blue lines represent calculations using US, with solid lines utilizing the CHARMM36 force field and dashed lines utilizing the CHARMM22/CMAP force field.

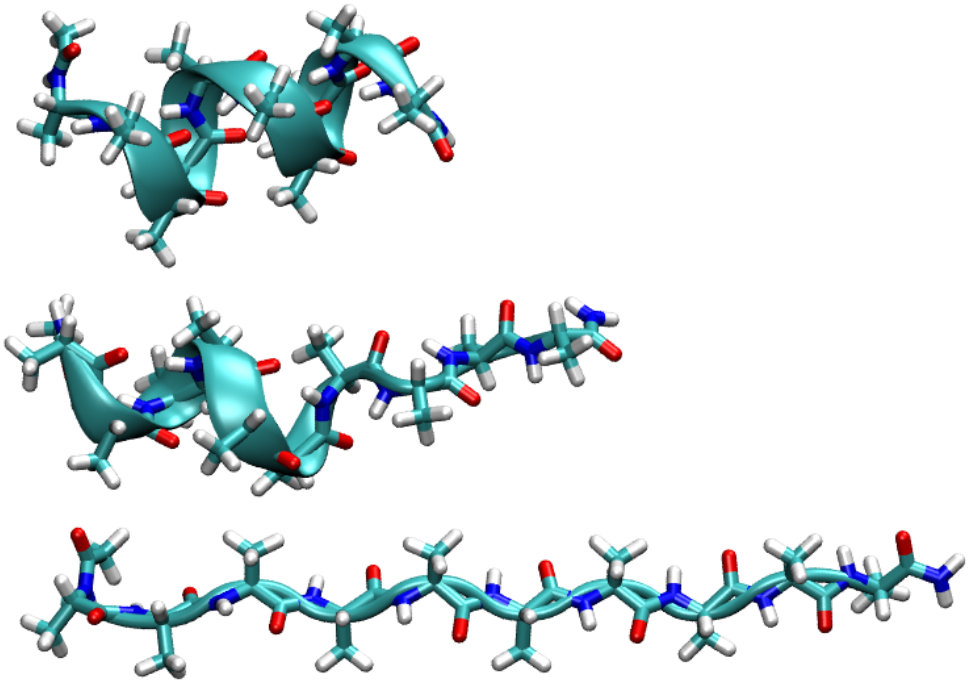


Figure 4.2: *Unfolding of Ala*$_{10}$. This "accordion-like" unfolding mechanism of Ala$_{10}$ was generated using SMD by pulling on the C-terminus C_α while keeping the N-terminus C_α fixed. Three snapshots of the peptide are shown in various stages of the SMD simulations, drawn using the "licorice" representation for all atoms and a cartoon representation for the backbone structure, where the thick ribbon represent those residues which are in an α-helical state. The top image represents the initial, minimized crystal structure of Ala$_{10}$ used as the starting state. The middle image represents an intermediate state in which the peptide is partially extended while the remaining portion of the peptide is still in an α-helical state. The bottom image represents the fully extended state.

PMFs [100, 101, 102, 103]. There is only one stable conformation around $\xi = 14.2$ Å, which corresponds to the pure α-helical state, and there is an energy barrier of $\sim$25 kcal/mol from the helical to extended state (Fig. 4.1). While the CHARMM22/CMAP force field does yield a minimum near the α-helical state, the entire PMF is shifted by $\sim$2Å toward lower end-to-end distances, and the energy barrier between helical and extended states is slightly higher ($\sim$30 kcal/mol). The corrections to the CHARMM22/CMAP force field added to the CHARMM36 force field are evident in the difference in the folding PMFs for Ala_{10}. Since CHARMM36 reproduces the expected free-energy minimum for the α-helical state in vacuum [104] and significantly improves agreement with helix-formation experiments [94], from here on we used solely the CHARMM36 force field.

The ABF and US simulations of Ala_{10} were then repeated in explicit water. Both methods yield a relatively flat PMF, compared to the vacuum PMF, along most of the reaction coordinate (Fig. 4.3, thick, solid lines). The trajectories reveal that there is no longer only the folding/refolding mechanism seen in vacuum; instead, the peptide transitions between extended states and compact, but non-helical, states. These non-helical states are characterized by various hairpin structures (Fig. 4.4). Fig. 4.5 shows the prevalence of low-helical, compact states in water as opposed to vacuum, which are "contaminating" the PMF at low end-to-end distances.

In order to enforce the α-helical folding/refolding mechanism observed in vacuum, multiple additional restraints were imposed. First, the peptide backbone was confined to a cylindrical tube of radius 10 Å centered along the end-to-end distance vector. This confinement, as well as a smaller tube of radius 8 Å, failed to prevent the formation of compact non-helical states, and the PMFs produced were either unchanged or inconsistent (Fig. 4.3, dashed lines), likely because convergence was not achieved. An additional anti-hairpin restraint, which prevents the backbone C_α's from passing one another in relation to the end-to-end distance vector, was also insufficient to produce the simple folding/refolding mechanism (Fig. 4.3, dotted-dashed lines). The persistent formation by Ala_{10} of these non-

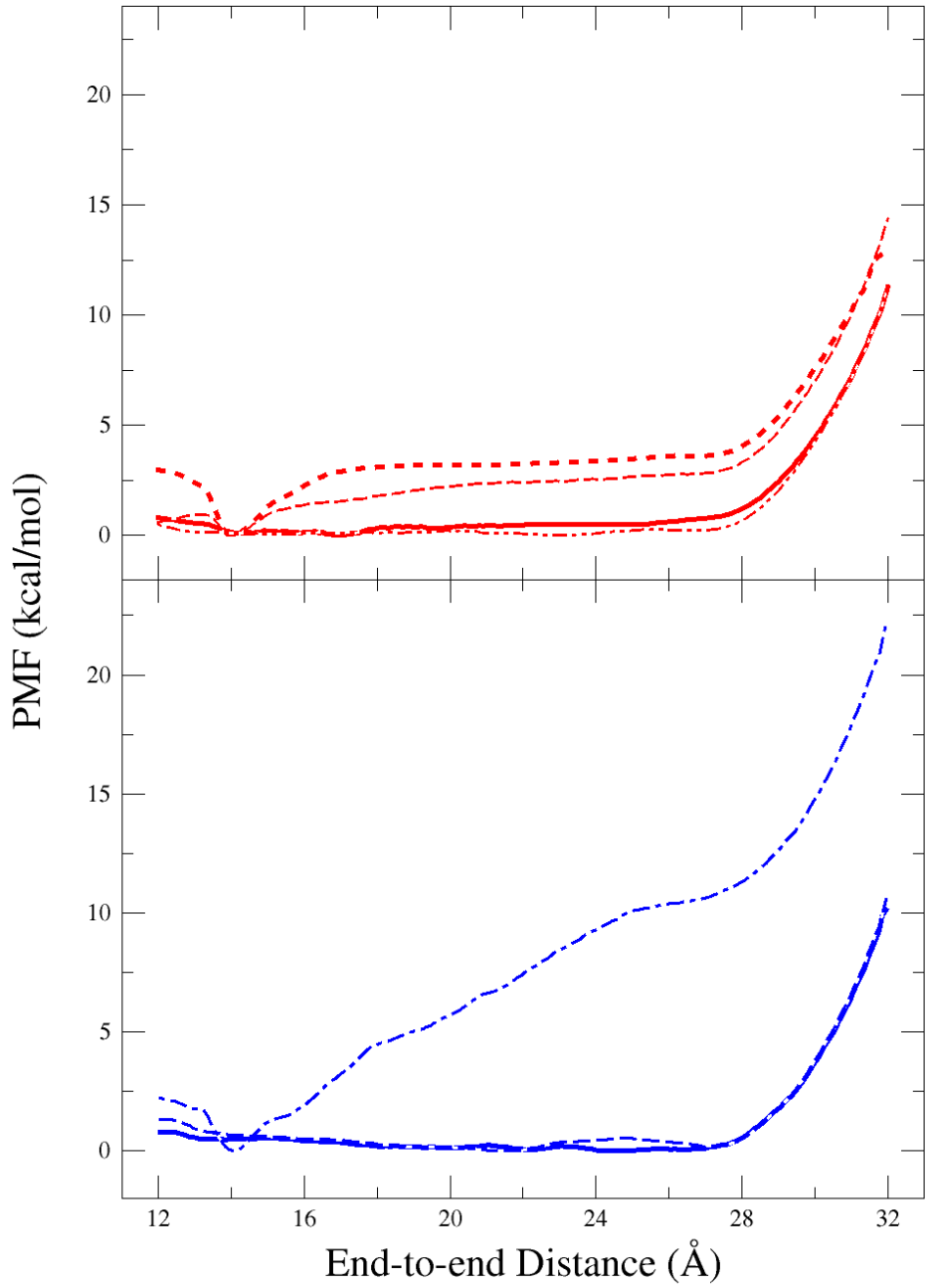


Figure 4.3: *One-dimensional PMF of Ala_{10} along the end-to-end distance of the peptide.* Calculated using ABF (top) and US (bottom). Top graph: no additional restraints (thick, solid line), no restraints with 50,000 full samples (thick, dashed line), 8 Å-radial confinement (thin, dashed line), 10 Å-radial confinement plus anti-hairpin restraint (thin, dotted-dashed line). Bottom graph: no additional restraints (thick, solid line), 10 Å-radial confinement (thin, dashed line), 10 Å-radial confinement plus anti-hairpin restraint (thin, dotted-dashed line).

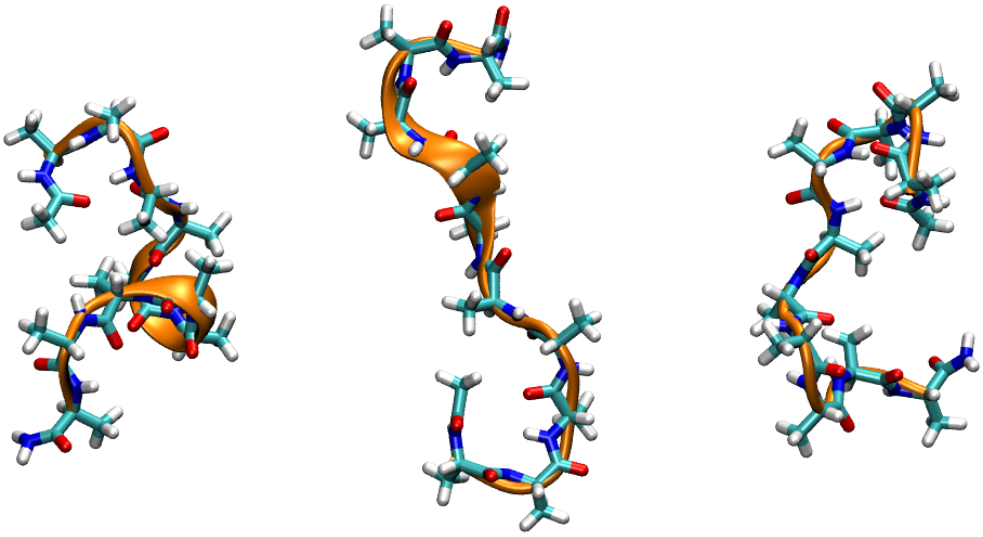


Figure 4.4: *Representative set of compact, low-α-helical content states of Ala$_{10}$ in water.* The Ala$_{10}$ peptide is shown in the "licorice" representation with the backbone α-helical content represented in orange by a cartoon representation. Water molecules are not shown.

helical, compact states from extended states reveals a more dynamic folding process than that seen in vacuum. Indeed, previous simulations of Ala$_{10}$ have shown that the disordered and extended states are much more soluble in water than the α-helix [105]. Instead of running the 1D US simulations longer in order to achieve convergence of the PMF, since the presence of compact, non-helical states makes convergence difficult to determine, we switched to a 2D description to ensure adequate sampling of these additional states.

4.3.2 Two-dimensional PMFs

To examine the effects of compact, non-helical states on the free energy of Ala$_{10}$ folding, we calculated a 2-dimensional PMF using US, with α-helical content as a second reaction coordinate. We first verified this 2D description by calculating the PMF in vacuum with umbrella sampling (Fig. 4.6, top). There is still only one minimum in the pure α-helical state, as was seen in the 1D PMF. In addition, we calculated the least free-energy path [106], which finds the path of least free-energy difference between two local minima on a 2D free-energy surface, from the α-helical state to an extended state. There is close agreement

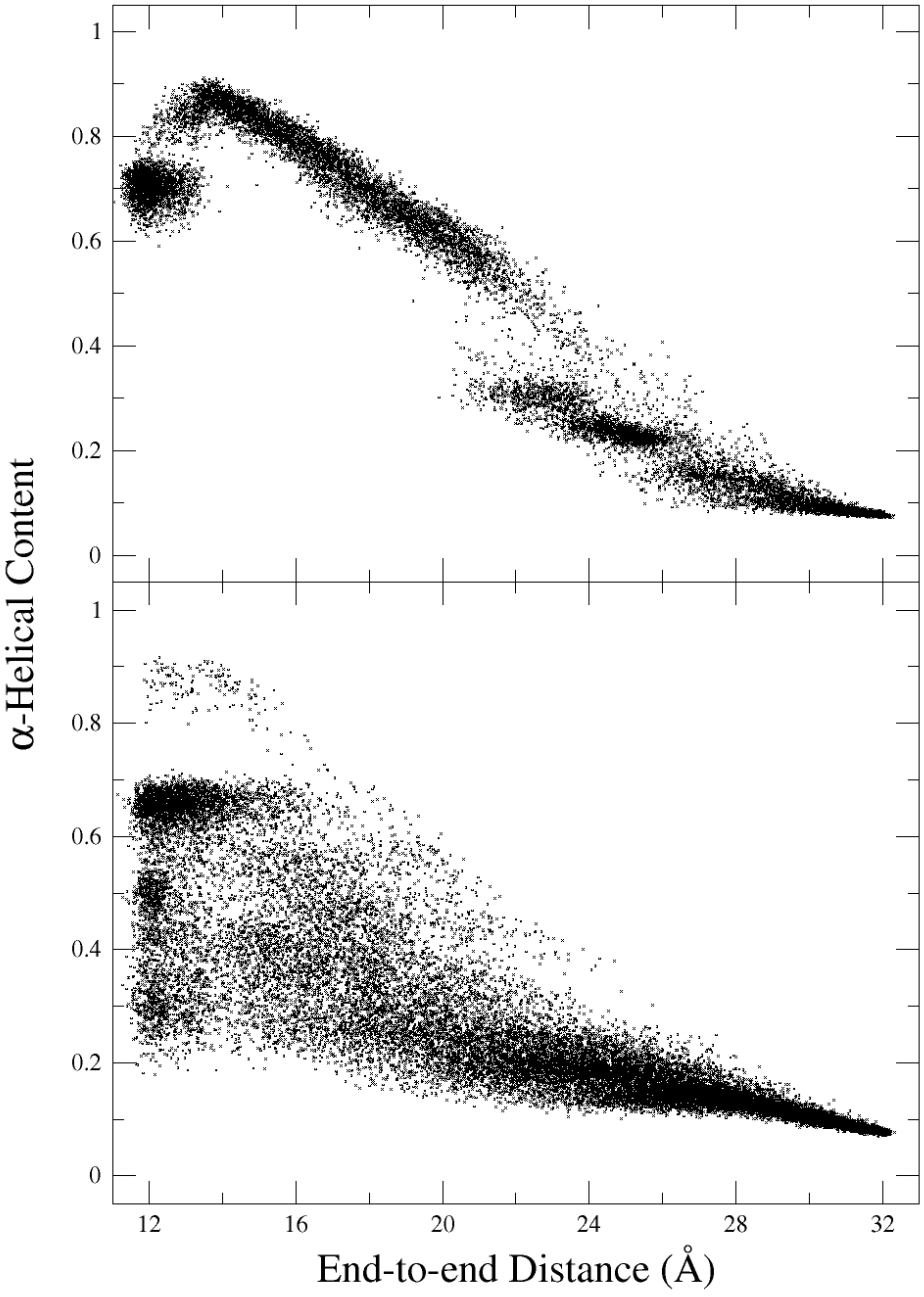


Figure 4.5: *Scatter plot of states from ABF simulations.* (Top) Scatter plot of states from 50 ns ABF simulation in vacuum using the CHARMM36 force field. (Bottom) Scatter plot of states from 100 ns ABF simulation in water with 10 Å-radial confinement plus anti-hairpin restraint.

between this path (Fig. 4.6, top inset) and the 1D PMF (Fig. 4.1), suggesting that the folding/refolding mechanism observed in the 1D biased simulations is in fact the primary mechanism of folding for Ala_{10} in vacuum.

Based on the successful application to the vacuum case, a 2-dimensional REMD-US (see Methods) simulation was performed for Ala_{10} in water. One can immediately see that a free-energy "trough" has appeared along a family of extended, non-helical states (Fig. 4.6, bottom). These states are also of free energy comparable to the pure α-helix, differing by less than 1 kcal/mol, and the energy barrier between the helical and extended states is now less than 4 kcal/mol. The least free-energy path explores a wider range of extended states before refolding compared with the vacuum case.

One notable feature of the 2D PMF is the appearance of "bands" in the free energy along lines of constant helical-content, which were presumed to be indications of poor overlap between neighboring windows when implementing the WHAM algorithm. The poor overlap was confirmed by plotting the histograms (data not shown) and, thus, the number of windows along the α coordinate was increased from 9 to 17 (see Methods). However, the bands still remained as seen in Fig. 4.6, bottom graph. These can be seen more explicitly in the PMF along the least free-energy path, which shows 5 distinct local minima between the α-helical state and the fully extended state (Fig. 4.6, bottom inset). There is less than 1 kcal/mol difference between PMFs calculated for 15ns per window and 20ns per window for the entire range of our reaction coordinates, which suggests that these PMFs should have reached convergence.

To validate the path as well as our choice of reaction coordinates, we also made a rough estimate of the committor distribution for the free-energy maximum [107, 108]. Fifty separation conformations were each used to initiate 50 10-ps simulations (2500 simulations and 25 ns in total) with random velocities. The committor was judged to be progressing to the extended state or retreating to the helical state based on the values of ξ and α (see Fig. 4.6, bottom) although full commitment to either minimum would require time scales

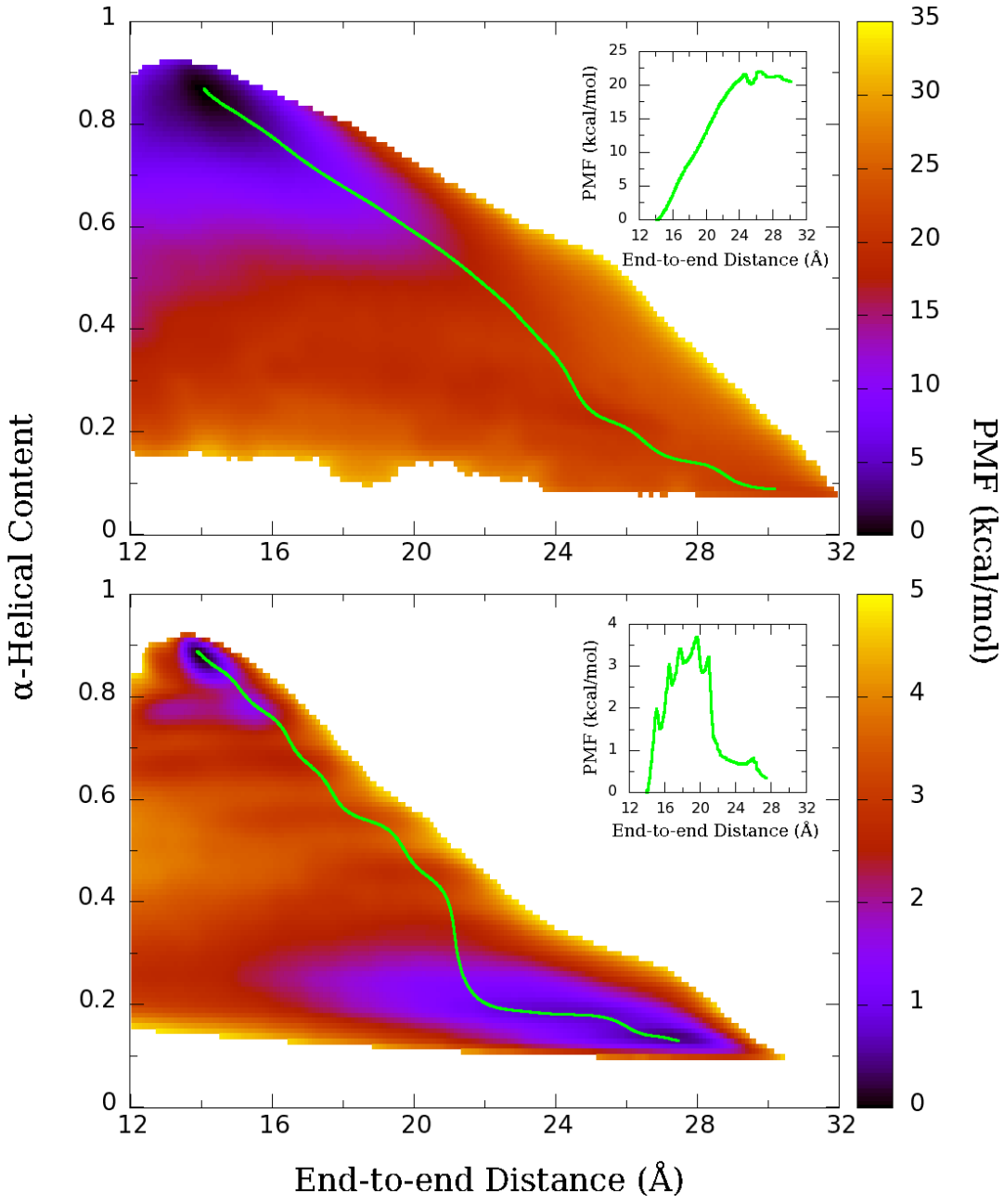


Figure 4.6: *Two-dimensional PMF of* Ala_{10}. Ala_{10} in (top) Vacuum and (bottom) water using end-to-end distance and α-helical content as the two reaction coordinates. Green line represents the least free-energy path from the α-helical state to the extended state. The inset shows the PMF along the least free-energy path, as projected onto the end-to-end distance coordinate.

at least 100× as long (see below). The resulting distribution is peaked at 0.5, with some bias towards values greater than 0.5 (see Fig. A.1). Overall, the behavior of the committor near the barrier suggests that it is representative of a true transition state.

4.3.3 Equilibrium Simulations

Equilibrium simulations of deca-alanine in water were performed starting from different states for 50 ns each to validate the 2D PMF and to better understand the folding pathway. Starting from a state that is near the α-helical minimum observed in the 2D PMF, in equilibrium the peptide initially samples the region around the minimum. As the simulation progresses, the peptide begins to unfold roughly along the least free-energy path and similar bands as seen in the PMF also appear in the histogram, despite no biasing being applied (Fig. 4.7). The protein folds and refolds until finally becoming fully extended near the end of the 50 ns simulation.

Starting from an extended state, Ala_{10} explores the range of extended and compact non-helical states predicted by the free-energy trough seen in the 2D PMF. The peptide eventually folds into an α-helix near the end of the simulation in much the same manner as in the unfolding case. Examination of the hydrogen bonding of Ala_{10} with itself and with water during the equilibrium simulations reveals that the transition between helical and extended states occurs in $\sim$5 ns with the formation or breaking of $\sim$4 peptide-peptide hydrogen bonds, with a corresponding decrease or increase in water-peptide hydrogen bonds, respectively (Fig. A.2), in agreement with the results of Ozer et al. [109]. We see similar results when examining the water-peptide hydrogen bonds from the REMD-US trajectories, with an increase of 8-10 hydrogen bonds from folded to extended states (Fig. A.3).

Based on the success of the two previous equilibrium simulations, we ran 20 additional simulations to get better sampling of the folding mechanism: 10 starting from the α-helical minimum, and 10 starting from the extended minimum. By examining the hydrogen bonding between the i–th residue and the $i + 4$–th residue for the simulations in which folding

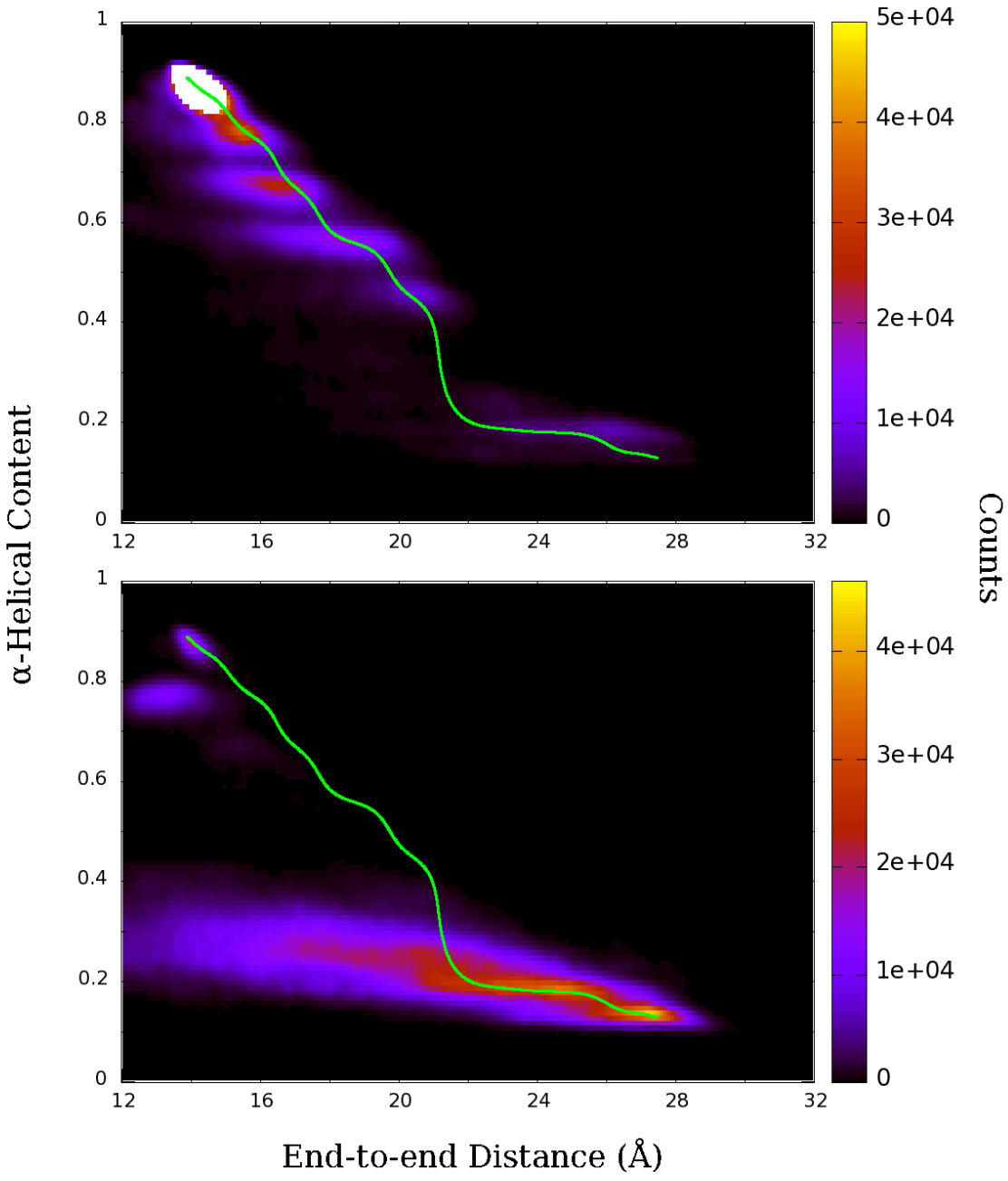


Figure 4.7: *Equilibrium simulations of Ala_{10} in water.* Histograms of 50 ns equilibrium simulations for (top) α-helical and (bottom) extended starting states. Green lines represent previously determined least free-energy paths.

was observed, some cooperativity is observed at the N-terminus, where the formation of the Ala1–Ala5 hydrogen bond initiates a propagation of hydrogen bond formation towards the C-terminus as the peptide folds from an extended state to an α-helical state (Fig. A.4, left graphs). In contrast the C-terminus exhibits less cooperativity, with unfolding typically beginning at the C-terminus and propagating in the opposite direction of folding in the majority of our simulations (Fig. A.4, right graphs). On average, the N-terminal hydrogen bonds persist longer than those on the C-terminal side while the protein is in a folded conformation. These results are consistent with a previous study which observed that the N-terminus of Ala_{10} slightly favors the α-helical conformation over the β-sheet conformation, whereas the opposite was observed for the C-terminus [64]. We do observe, however, cases in which the N-terminal hydrogen bonds were broken while the C-terminal hydrogen bonds remained intact (Fig. A.4, bottom right graph), although these occurred much less frequently. The REMD-US trajectories of Ala_{10} in water yields similar preference for N-terminal hydrogen bond formation (Fig. A.5) near the α-helical state. The N-terminal hydrogen bonds also persist for a longer portion of the unfolding pathway (Fig. 4.6, bottom graph) than the C-terminal hydrogen bonds. In addition, we observed very little 3_{10}-helical $(i, i+3)$ hydrogen bonding for both the REMD-US and equilibrium trajectories (data not shown), so it appears that the 3_{10}-helix is not an intermediate state in the folding pathway. Thus, the folding pathway consists of the breaking or formation of α-helical hydrogen bonds and not the rearrangement of those hydrogen bonds into a 3_{10}-helical structure.

4.3.4 1D PMF from Integration of 2D PMF

After validation of the free energy minima observed in the 2D PMF of Ala_{10} folding in water by equilibrium simulations, we integrated out the α coordinate according to Eq. 4.1 (see Methods) to generate a 1D PMF along the distance coordinate (Fig. 4.8). This integrated PMF still yields the free-energy minimum observed for Ala_{10} in vacuum around $\xi = 14.3\,\text{Å}$. The main difference between the integrated PMF in Fig. 4.8 and the PMF in

vacuum (Fig. 4.1) is in the unfolding region ($\xi > 15\,\text{Å}$), with the PMF reduced by more than 20 kcal/mol in the extended state. This reduction is comparable to that seen in the previous unrestrained 1D PMFs calculated for Ala_{10} (see Fig. 4.3). However, by ensuring that Ala_{10} more fully explores its entire conformational space through biasing of the additional α reaction coordinate, two free energy minima are revealed in the compact states and extended states, respectively, that were not found by biasing of the ξ reaction coordinate alone. The appearance of these new minima supports the suggestion that the previous 1D PMFs had not yet converged. Calculation of the free-energy difference between these two minima establishes that the compact state is slightly favored over the extended states ($\Delta G = -0.4$ kcal/mol), with compact states defined as $\xi \leq 16.75\,\text{Å}$, i.e., below the peak of the energy barrier between the minima.

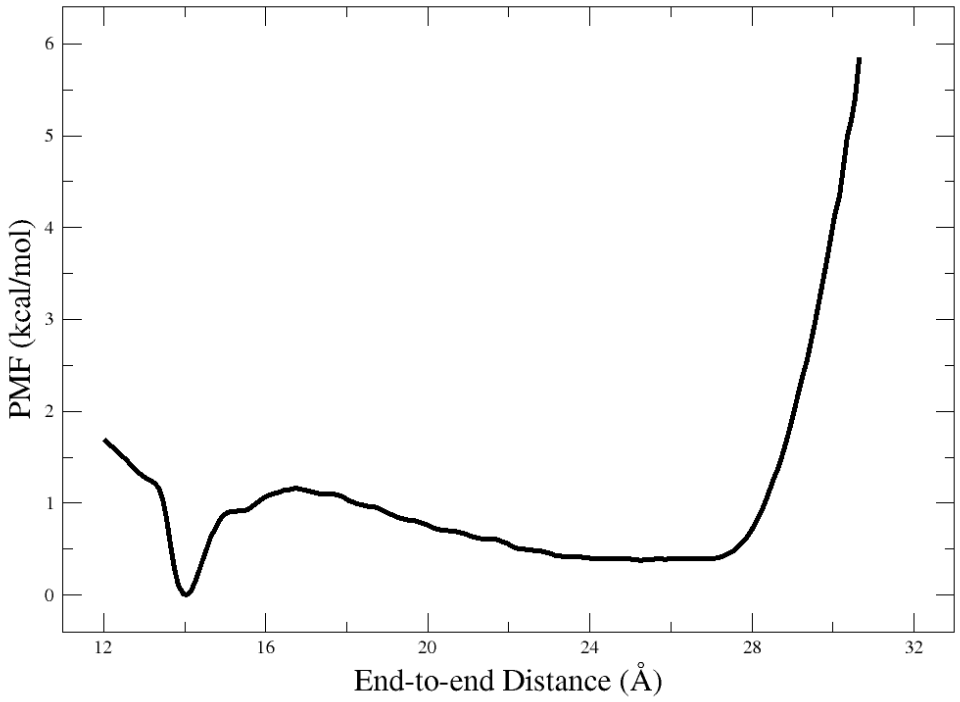


Figure 4.8: *One-dimensional PMF of Ala_{10} in water.* Calculated by integration of 2D PMF using Eq. 4.1.

4.4 Discussion

The free energy of folding for Ala_{10} in vacuum has been used as a benchmark for many free-energy calculation methods, using the end-to-end distance (ξ) of the peptide as the reaction coordinate. This choice of reaction coordinate implicitly presumes a reversible accordion-like folding/refolding of the peptide. We observed using both US and ABF that in vacuum, this presumption is indeed correct and only the simple folding/unfolding mechanism predicted is found. In the presence of water, however, the folding/unfolding mechanism is much more complex, and biasing the refolding of Ala_{10} back into an α-helix from an extended state is non-trivial.

Using multiple biasing methods – US and ABF – we discovered that the water-solvated Ala_{10} will explore an extended range of compact, non-helical states before refolding back into an α-helix. These compact, non-helical states appeared to be "contaminating" the low-ξ region of the 1D PMFs calculated from the two biasing methods creating a relatively flat PMF compared to the PMF calculated for Ala_{10} in vacuum (Fig. 4.3). Calculation of 2D PMFs for the entire (ξ, α) collective-variable space revealed a new free-energy minimum in a family of extended states not observed in vacuum, along with the α-helical minimum that was originally observed in vacuum.

In water, the free energy of the extended states decreased significantly compared to vacuum, with the α-helical state less than 1 kcal/mol lower in energy than the extended states – a decrease from more than 20 kcal/mol observed in vacuum. A barrier of $\sim$4 kcal/mol between the two energy minima is also observed in the PMF. Previous studies of Ala_{10} in water have also shown a decrease in the free-energy difference between extended and helical states. For example, ABF was applied to a zwitterionic form of Ala_{10} with charged termini to calculate a 1D PMF in water, using the average length of the $i, i+4$ hydrogen bonds of the backbone as a reaction coordinate [35]. That PMF shows a comparable free-energy barrier of $\sim$5 kcal/mol between the extended and helical states, with the relative

energies differing by $\sim$1 kcal/mol. The decrease in the free-energy difference was not as pronounced in other studies utilizing Adaptive SMD [109] and US [110] applied to the end-to-end distance of Ala_{10} with neutral termini. Additionally, neither study discovered the free-energy minimum in the extended states. Levy et al. found that in hydrophilic environments, the α-helix is actually destabilized relative to β-sheet conformations, due to their high conformational entropy compared to that of the α-helix [104]. Although we observed a minimum in extended states rather than β-sheet conformations, the extended states are similarly entropically favored over the α-helical state and extend into the range of compact, non-helical states (Fig. 4.6). As a check on our 2D PMF, we performed 50 ns equilibrium simulations of Ala_{10} in water starting from α-helical and extended states (Fig. 4.7). These simulations confirmed both minima observed in the 2D PMF. Furthermore, transitions between the two minima demonstrated cooperativity at the N-terminus and non-cooperativity at the C-terminus, as expected [64].

Although alanine is used as a model for protein folding since it has the highest helix propensity of all amino acids, the folding mechanism for Ala-based peptides is still not very well understood. Experiments studying short Ala-based peptides have lead to inconclusive or contradictory observations for the stability of the α-helical state in water. Rohl et al. observed that Ala-based peptides are the only stable helix formers in water for the 20 common amino acids [111]. They postulated that reducing the extent of solvation of the coil backbone could increase stabilization of the helix. Experiments performed by Blondelle et al. observed that peptides of the form Ac–KYA_nK–NH_2 ($10 \leq n \leq 14$) coexisted as a β-sheet and α-helix to varying extents [112]. However, it was later observed that the stability of the helix in these peptides was due to the solubility of the flanking Lys residues, and not the intrinsic helix propensity of Ala [113]. This was followed up by Spek et al. stating that although the increase in α-content of KA_nK is an artifact of the flanking Lys residues, Ala is intrinsically α-helix stabilizing [114]. So, although there is some discrepancy for the stability of secondary structures for Ala-based peptides, the evidence suggests

that these peptides do not solely exist as an α-helix in solution as one might suspect in vacuum. Instead they exist in some combination of secondary structures, including α-helices and β-sheets. Indeed, more recently, NMR data for short polyalanine peptides (Ala_{3-7}) shows that these peptides exist primarily as polyproline II (PP_{II}) helix-like structures with very little population of the α-helix [115]. It is also known that macromolecular crowding has a significant effect on protein folding *in vivo* [116, 117]. One could study this effect by introducing multiple Ala_{10} molecules or other macromolecules, but that is beyond the scope of this work.

Our results could shed some light on the stability of the α-helix for short polyalanine peptides in solution. However, one should always be skeptical of the force fields utilized in MD simulations. Best et al. examined a range of force fields and observed that they overemphasized the α-helix structure compared with NMR coupling parameters for the backbone (ϕ, ψ) dihedrals [118]. By re-weighting the force fields based on these (ϕ, ψ) coupling parameters [115], they were able to yield good agreement with the population of the α-helix, β-sheet, and PP_{II} structures seen in the NMR data. Although there was better agreement with NMR for unblocked (ionic or zwitterionic) termini than with blocked (neutral) termini, by using the re-weighting for unblocked peptides, blocked peptides were found to yield α-content of 12–23%. Similar results were found using Agadir [119, 120, 121, 122, 123], an empirical NMR-based algorithm that determines the α-helical propensity of a peptide based on sequence, which yields a helical propensity of $\sim$15% for Ac–Ala_{10}–NH_2 in water at 300 K.

In this study we have used the most recently updated version of the CHARMM force field, CHARMM36 [94, 95]. One of the major improvements of CHARMM36 is the correction of the α-helical bias introduced into CHARMM22 by the backbone (ϕ, ψ) dihedral CMAP potential. This improvement was achieved by capturing the many-body effects not present in the original CMAP potential. The CMAP potential was optimized to match NMR data for Ala_5 [115] and Ac-$(AAQAA)_3$-NH_2 [124] in solution, and the side-chain χ_1

and χ_2 dihedrals were optimized to QM energy surfaces. The result was a better balance between secondary structures, particularly the α-helix and β-sheet, addressing the problem posed by Best et al. [118]. Particularly, the fraction helix of Ac-$(AAQAA)_3$-NH_2 produced by CHARMM36 more closely matches experiments from Schalongo et al. [124] than other force fields (a reduction from 95% for CHARMM22/CMAP to 21% for CHARMM36), as well as improved cooperativity for α-helix and β-sheet formation. Thus, the fraction helix of Ala_{10}, as well as its folding mechanism, determined using CHARMM36 should also have better agreement with experiment, particularly when compared with CHARMM22 and CHARMM22/CMAP, which were utilized in previous unfolding simulations of Ala_{10} in water [109, 110].

Our work emphasizes the challenge and necessity of choosing relevant reaction coordinates to fully characterize a particular system [125]. For Ala_{10}, the end-to-end distance is no longer sufficient to capture the diversity of conformations explored in water, thus making it a highly degenerate reaction coordinate (Fig. 4.4). Contributions from compact, non-helical states can produce a PMF that doesn't reveal what one intuitively expects it to, namely the accordion-like folding/refolding mechanism shown in Fig. 4.2. By tracking the α-helical content of Ala_{10} during biased folding simulations, we found many states accessible to the peptide in water that were inaccessible in vacuum. These states arise due to Ala_{10}'s increased flexibility in water, where a loss of intrapeptide hydrogen bonds is compensated by an increase in peptide-water hydrogen bonds (Figs. A.2 and A.3). Reaction coordinates suitable for Ala_{10} in vacuum, therefore, may not be suitable in water [46, 104]. Since recent studies of Ala_{10} folding in water have only used the end-to-end distance to characterize the folding pathway, they observe that the preference for the α-helical state is still significant in water compared to the extended states. However, we have shown that the helical and extended states are of comparable stability ($|\Delta G| < 1$ kcal/mol), with both states transitioning from one to the other within the course of 50 ns equilibrium simulations.

CHAPTER 5

FOLDING FREE ENERGY LANDSCAPES OF β-SHEETS WITH NON-POLARIZABLE AND POLARIZABLE CHARMM FORCE FIELDS

5.1 Introduction

The so-called "protein folding problem", namely how proteins arrive at a particular three-dimensional shape determined primarily by their sequence of amino acids, remains one of the grand challenges of molecular biology [126]. While most estimates of the number of distinct protein folds are around 10^3 (see Ref. [127]), they practically all derive from the two most common secondary-structure elements: the α-helix and the β-sheet. Thus, model systems exemplifying one of these elements represent important tools for the study of protein folding.

One of the most common representatives of the β-sheet used in protein folding studies is the B1 domain of streptococcal protein G (GB1) as well as its C-terminal β-hairpin (residues 41-56), which comprises just 16 amino acids. Although other isolated β-hairpin peptides have been found [128, 129], GB1 has proven particularly amenable to experiments looking at folding dynamics and mechanisms [122]. GB1 is stabilized by six native backbone hydrogen bonds and a hydrophobic core consisting of four residues (see Fig. 5.1). Muñoz et al. measured folding to be a two-state process taking $\sim$6 μs, and multiple studies have found the fraction folded to be 30-50% at room temperature [122, 130]. This folding time is $\sim$30x slower than that for α-helices [63, 131]. The earliest folding model put forward proposed that the β-turn forms first and then the hairpin "zips up" from there, with the hydrophobic core forming last [122, 132].

Because of its small size and fundamental nature, the folding of the GB1 β-hairpin has inspired a number of molecular dynamics (MD) studies. In particular, multiple groups have

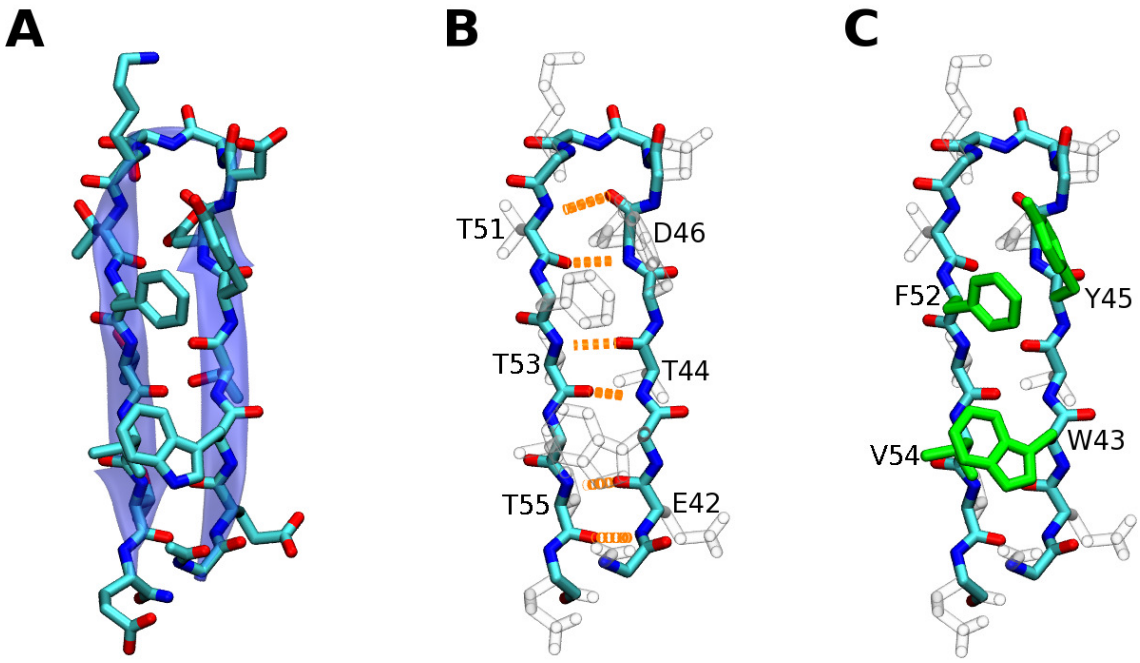


Figure 5.1: *Crystal structure of the C-terminal β-hairpin of GB1.* Residues 41-56 of the full GB1 crystal structure (PDB: 1GB1). All protein atoms are portrayed in the licorice representation. (A) All non-hydrogen atoms, colored by atom name. β-Hairpin structure shown in cartoon representation in transparent blue. (B) Internal, backbone hydrogen bonds are represented by orange dashed lines. Residue pairs 46–51, 44–53, and 42–55 are labeled. Side chains are shown as transparent. (C) Hydrophobic core (residues 43, 45, 52, and 54) side chains are colored green. All other side chains are shown as transparent.

calculated the free-energy landscape of GB1. For example, Dinner et al. chose two reaction coordinates to map GB1's conformations, namely the native hydrogen bonds and the packing of the hydrophobic core residues; they found, contrary to the earlier model, that the hydrophobic core forms first, followed by hydrogen bond formation along the backbone [133]. Successive simulation studies agreed with this model [134, 135]. Zhou and Berne compared explicit and implicit solvent models with the same force field, finding that the folded probability of GB1 in the implicit solvent was much lower than that in the explicit solvent, the latter matching experiments [136]. More recently, Shao et al. demonstrated that a particular implicit solvent model, GB^{OBC}, when used with the AMBER ff96 force field, could reproduce the expected structures of GB1 [137]. Thus, GB1 β-hairpin folding developed into a useful and sensitive test of simulation force fields. Best and Mittal further demonstrated this utility by comparing six force-field/explicit-water-model combinations; they found that the primary differences between force fields are the residual secondary structures observed in the unfolded state as well as the dimensions of the unfolded state, which was often too compact [138]. The delicate balance between solute-solvent and solute-solute interactions was cited as a reason for the differences.

Force fields that include explicit polarizability have grown in popularity over the years, due in part to improvements in the methods and also the expansion in computational resources available [139, 140, 141, 142]. Two commonly used approaches that have been extended to proteins are the AMOEBA force field [29] and the Drude-2013 force field [33]. The AMOEBA-2013 force field proved adept at maintaining conformations of ten small proteins with varied folds (α-helices, β-sheets, and mixed) over 30-ns simulations [29]. The Drude-2013 force field was used for simulations of a few small peptides as well as ten proteins, finding them to be stable on the 100-ns time scale [33]. Dynamic polarization of the peptide backbone led to improvements in both the temperature dependence and cooperativity of α-helix folding for the $(AAQAA)_3$ peptide [143]. Lopes et al. also simulated the GB1 β-hairpin with Drude-2013. While it remained close to the crystal structure

over the course of a 150-ns simulation, comparison of estimated NMR chemical shifts with experiment for residues 42-45 proved worse than the non-polarizable CHARMM36 force field [33]. Additionally, backbone hydrogen bonds in β-sheet structures were approximately 0.05 Å too long compared to high-resolution crystal structures [33].

In the present study, we determine the free energy of GB1's conformational landscape for three force fields: CHARMM36 (Ref. [95]), CHARMM22* (Ref. [144]), and the 2015 release of the Drude-2013 model (Ref. [33]). We find that while the β-hairpin of GB1 is stable in the non-polarizable models, it is unstable in the Drude-2013 model, with the latter favoring an unfolded, extended state. Hydration free energies reveal that amide-containing groups are more soluble in the Drude-2013 model than in the non-polarizable models, resulting in a more soluble backbone. However, this enhanced solubility, which is still less than found in experiments, does not appear to be countered by a corresponding enhancement in peptide–peptide interactions, namely for the native hydrogen bonds that form the β-hairpin. Using a perturbative approach, we show that small increases to the backbone amide N polarizability can improve both the backbone solubility and folding properties of the GB1 β-hairpin in the Drude-2013 model without sacrificing the improvements observed for α-helical peptides.

5.2 Methods

5.2.1 Non-polarizable model

All-atom simulations of non-polarizable GB1 were performed starting from residues 41–56 of PDB 1GB1 (Ref. [145]; see Fig. 5.1), capped with an acetylated N-terminus and amidated C-terminus. For non-polarizable models, the visualization and analysis program VMD [91] was used to place GB1 in a water box of 14,314 TIP3P (Ref. [92]) water molecules with dimensions $66 \times 66 \times 104$ Å^3. Systems were neutralized with 6 sodium and 3 chloride ions as used previously [138], for a total of 43,204 atoms. A harmonic restraint was added to keep the end-to-end vector along the z-axis. Molecular dynamics

simulations were carried out using NAMD 2.10-12 (Ref. [**Phillips2005**]) with CHARMM all-atom force fields C22* (Ref. [144]) and C36 (Refs. [94, 95]).

The temperature was fixed at 300 K using Langevin dynamics; the pressure was kept constant at 1 atm using the Langevin piston method [96]. The equations of motion were integrated using the RESPA multiple time-step algorithm with a time step of 2 fs used for all bonded interactions, 2 fs for short-range non-bonded interactions, and 4 fs for long-range electrostatic interactions. Long-range electrostatic interactions were calculated using the particle-mesh Ewald method with a real-space cutoff of 12 Å [97]. Short-range non-bonded Lennard-Jones interactions were cutoff at 12 Å with a potential switching function beginning at 10 Å bringing the potential energy smoothly to zero at the cut off distance. Bonds involving hydrogen atoms were constrained to their equilibrium length, employing the SETTLE algorithm [146] for water molecules and the SHAKE algorithm for all others [147].

5.2.2 Polarizable model

For the polarizable model, a non-polarizable model was first built using CHARMM-GUI [148]. Similar to the previous non-polarizable models, GB1 was solvated in a water box of dimensions $66 \times 66 \times 104$ Å^3 with 14,256 water molecules and neutralized with 6 sodium and 3 chloride ions. An additional harmonic restraint was added to keep the end-to-end vector along the z-axis. The Drude Prepper [149] from CHARMM-GUI was then used to convert the non-polarizable model into the Drude-2013 polarizable model for the protein [33] and ions [150] solvated in SWM4-NDP polarizable water molecules [151], resulting in a system of 71,748 atoms, including Drude particles. We used the Drude-2013 parameters released in July 2015. The system was then minimized and pre-equilibrated using the NAMD input scripts provided by Drude Prepper.

The temperature for parent atoms was fixed at 300 K using Langevin dynamics; a separate Langevin thermostat was coupled to the Drude particles with temperature 1 K. Addi-

tionally, a hard wall restraint at 0.25 Å was added for all Drude particle-parent atom bond lengths. Thole corrections to electrostatic interactions [152] were also extended to non-bonded pairs of Drude oscillators for which a Thole screening parameter is defined within a 5.0-Å cutoff. Analytic long-range Lennard-Jones corrections were also implemented [153]. The pressure was kept constant at 1 atm using the Langevin piston method [96]. The equations of motion were integrated using a 1-fs time step for all interactions. Long-range electrostatic interactions were calculated using the particle-mesh Ewald method [97]; cutoffs were the same as for the non-polarizable simulations above. Bonds involving hydrogen atoms were constrained only for water molecules.

5.2.3 Folding free energy calculations

Two-dimensional potentials of mean force (PMF) were calculated using umbrella sampling with replica-exchange (REUS) [51]. Reaction coordinates were calculated and biased using the *collective variables* (colvars) module of NAMD 2.10 (Ref. [154]). The first reaction coordinate is the number of native hydrogen bonds (N_{hb}), calculated using the `hbonds` colvar. N_{hb} is the sum of individual dimensionless hydrogen bond scoring functions, hbf, calculated for each donor and acceptor pair:

$$\mathrm{hbf}(\mathrm{donor}, \mathrm{acceptor}) = \frac{1 - (|\boldsymbol{x}_{\mathrm{don}} - \boldsymbol{x}_{\mathrm{acc}}|/d_0)^6}{1 - (|\boldsymbol{x}_{\mathrm{don}} - \boldsymbol{x}_{\mathrm{acc}}|/d_0)^8}, \tag{5.1}$$

where $\boldsymbol{x}_i$ is the cartesian coordinates of atom i, and $d_0 = 3.3$ Å is the distance cutoff for an occupied hydrogen bond. $\mathrm{hbf} \in [0, 1]$, with values near 0 for donors and acceptors far outside the cutoff distance, and values near 1 well inside the cutoff. Therefore, $N_{\mathrm{hb}} \in [0, 6]$ for the six native hydrogen bonds of GB1. The second reaction coordinate is the radius of gyration of the hydrophobic core (R_{G}), i.e., the non-hydrogen atoms of residues Trp43, Tyr45, Phe52 and Val54. Either 89 (C36 and Drude-2013) or 100 (C22*) windows were simulated for 15 ns/window, for a total simulation time of 1.34-1.50 μs per system.

Windows were spaced along the N_{hb} coordinate by 0.25 for $N_{\mathrm{hb}} \leq 1.5$ and by 0.5 for $N_{\mathrm{hb}} > 1.5$, using a 25 kcal/mol harmonic force constant. Windows were spaced along the R_{G} coordinate by 1.0 Å, using a 6.25 kcal/mol·Å^2 harmonic force constant. The first 2 ns/window of each system was omitted before calculating the PMF, generated via the weighted histogram analysis method (WHAM) [45, 155]. Starting states for each window were generated using steered molecular dynamics (SMD) [156].

1D PMFs were calculated by integrating out the second coordinate as follows [99]:

$$e^{-\beta w(x)} = e^{-\beta W(x_c,y_c)} \frac{\int e^{-\beta W(x,y)} dy}{\int e^{-\beta W(x_c,y)} dy}, \tag{5.2}$$

where $W(x, y)$ is the 2D PMF, $w(x)$ is the corresponding 1D PMF, and (x_c, y_c) is an arbitrary point in the collective variable space. To calculate folding free energies, ΔG_{fold}, from the 1D PMFs, we first located the peak of the PMF between the two extrema of the x-coordinate; then, for folding-related coordinates (e.g. number of hydrogen bonds or fraction of native structure), the 1D PMF was integrated to the left and to the right of the peak to get G_{unfolded} and G_{folded}, respectively, where $\Delta G_{\mathrm{fold}} = G_{\mathrm{folded}} - G_{\mathrm{unfolded}}$. For distance-related coordinates (e.g., radius of gyration or end-to-end distance), the definitions for folded and unfolded are reversed. Errors for ΔG_{fold} were calculated as the standard deviation of ΔG_{fold}s calculated from 1-ns/window blocks of the REUS trajectories.

5.2.4 Adjusting Drude polarizability and charge parameters

Drude particle charges, q_{D}, and parent atom charges, q_{P}, are determined by

$$\begin{aligned} q_{\mathrm{D}} &= (2K_{\mathrm{D}}\alpha)\mathrm{sgn}(\alpha)\mathrm{and} \\ q_{\mathrm{P}} &= q_{\mathrm{A}} - q_{\mathrm{D}}, \end{aligned} \tag{5.3}$$

where q_{A} is the atomic partial charge, $K_{\mathrm{D}} = 500$ kcal/mol·Å^2 is the strength of the harmonic potential connecting the Drude particle to its parent atom, α is the atomic polariz-

ability, and sgn(α) = $+1$ if $\alpha > 0$, -1 if $\alpha < 0$, and 0 if $\alpha = 0$. Parent charges also include any associated lone pair particles. If we modify the polarizability by a factor $\alpha'/\alpha = \gamma$, then $q'_{\mathrm{D}} = \gamma^{1/2} q_{\mathrm{D}}$ and $q'_{\mathrm{P}} = q_{\mathrm{A}} - q'_{\mathrm{D}}$. For atoms with lone pair particles, we keep the parent charge constant and split the new charge between the lone pair particles.

States generated from the REUS simulations were used to generate new PMFs perturbatively. First, the total potential energy was calculated for 10-ps snapshots from the last 5 ns/window of the REUS simulations. Then, new potential energies were calculated for these same snapshots using adjusted parameters. WHAM histograms were reweighted by the difference in the energies using Eq. 5.4 below, and new PMFs were calculated:

$$\mathrm{histogram}(r_{ij}) = \frac{\sum_t \exp\big(-\beta(U_{\mathrm{new}} - U_{\mathrm{old}})\big)\delta(r(t) - r_{ij})}{\sum_t \exp\big(-\beta(U_{\mathrm{new}} - U_{\mathrm{old}})\big)}, \tag{5.4}$$

where U_{new} and U_{old} refer to the potential energy $U(r(t))$ with the new and old parameters, respectively; $r(t)$ is the value of the collective variables at timestep t; and r_{ij} is the value of the collective variables in bin ij. For consistency, the original PMFs were also recalculated using the same 10-ps snapshots to account for any bias in the reduced data set.

5.2.5 Hydration free energy calculations

The sidechain hydration energies were calculated using CHARMM c40b1 [157]. The 40-stage WCA-decomposition FEP procedure described by Deng and Roux was used [158, 159]. The electrostatic and dispersion components were calculated from windows from $\lambda = 0.0, 0.1, 0.2, ..., 1.0$. Exchanges were performed between neighboring replicas with a frequency of 2 ps. The solutes were placed in a $32 \times 32 \times 32$ Å simulation cell of TIP3P-model water (roughly 980 water molecules). The Gibbs energy calculations were performed using a 2 ns simulation for equilibration followed by a 5 ns simulation to sample the free energies. Gibbs energies were calculated using WHAM [45]. For charged amino acids, the electrostatic component solvation energies were corrected for an interfa-

cial potential of -520 mV [160]. Uncertainties were estimated by calculating the standard deviation of the free energies taken from thirds of the production simulation.

A correction for the neglect of dispersion interactions outside the 12-Å cutoff distance was determined by calculating the average of the difference between the interaction energy with a 12-Å Lennard-Jones cutoff and that with a 45-Å Lennard-Jones cutoff in a 1-ns MD simulation. This energy is included in the dispersion component of the hydration energy.

The Drude-2013 model hydration energies were calculated using the same procedure; however, a 1-fs time step was used. The SWM4-NDP water model was used as the solvent. The electrostatic components of the FEP of the Drude-2013 hydration energies were determined without exchanges. An interfacial potential of -545 mV was used to calculate the correction to the electrostatic component of the solvation energy for ionic solutes [161].

5.3 Results and discussion

5.3.1 GB1 PMFs

Two-dimensional potentials of mean force (PMF) were calculated for GB1 using CHARMM36 (C36), CHARMM22* (C22*), and CHARMM Drude Polarizable (Drude-2013) force fields (see Fig. 5.2) using replica exchange umbrella sampling (REUS, see Methods). Previous simulation studies of GB1 near room temperature using a variety of protein force fields and solvent models have found a number of local free energy minima, including the fully folded state, partially folded states, compact unfolded states, and extended states [134, 136, 137, 138, 162, 163, 164, 165, 166]. We find that the three CHARMM force fields also produce a variety of local minima, resulting in distinct folding energetics for each model. The free energy landscape for C36 shows the native β-hairpin ($\sim$4-5 native hydrogen bonds) to be very stable, roughly 6 kcal/mol lower in energy than the compact, unfolded state and 8 kcal/mol lower than the extended state. The folded state is much less stable in C22*, with a free energy basin composed of multiple local minima extending from the native hairpin with 5 hydrogen bonds to a partially unfolded hairpin with 2 hydrogen bonds, with the low-

est energy state residing near 3 hydrogen bonds; local minima are all separated by barriers of <2 kcal/mol. This free energy basin is around 2-4 kcal/mol lower in energy than both the compact and extended unfolded states. While C36 and C22* have global minima in the native and partially folded states at coordinates (N_{hb}, R_{G}) = (4.95, 5.45) and (3.11, 5.45), respectively, the Drude-2013 model shows a global minimum in the extended, unfolded state at coordinates (0.23, 11.65), with shallower local minima in the partially folded and fully folded regions.

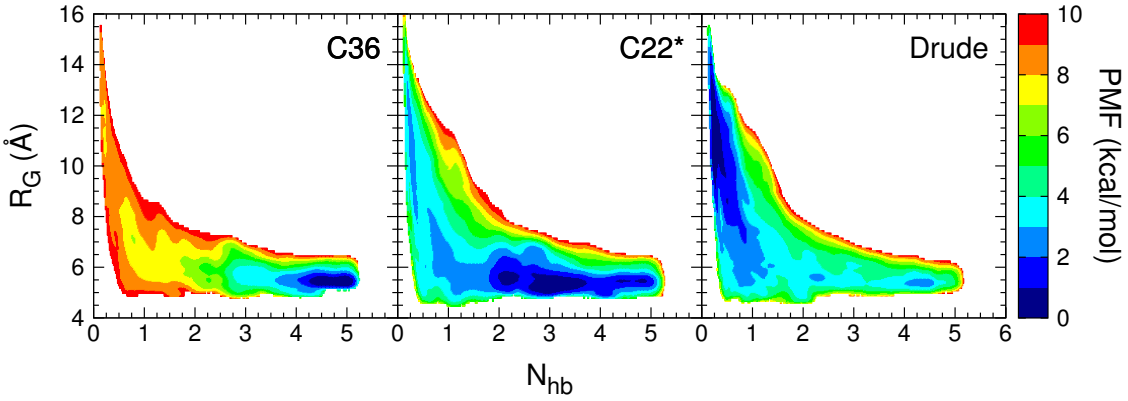


Figure 5.2: *Folding free energy landscapes of GB1.* 2D PMFs calculated by REUS using (left) C36, (middle) C22*, and (right) Drude-2013 models. A total of 89 (C36/Drude-2013 systems) or 100 (C22* system) windows were utilized and simulated for 15 ns/window. The first 2 ns/window for each system was omitted before calculating PMFs.

To calculate the folding energetics, we integrated out the N_{hb} and R_{G} coordinates separately from our 2D PMFs to generate 1D PMFs along the other corresponding coordinate (see Fig. B.1) and calculated the free energy of folding, i.e. $\Delta G_{\mathrm{fold}} = G_{\mathrm{folded}} - G_{\mathrm{unfolded}}$ (see Methods). The results are shown in Table 5.1. From the 2D PMFs, the N_{hb} coordinate appears to better differentiate between the folded and unfolded states than the R_{G} coordinate, with a clear barrier between the two states for all three force fields. Hereafter, we will only compare folding free energies along the N_{hb} coordinate. GB1 is between 30 and 50% folded at room temperature [122, 130], so ΔG_{fold} = 0.0-0.5 kcal/mol. C36 overstabilizes the folded state by ∼8 kcal/mol, C22* overstabilizes it by ∼2 kcal/mol, and Drude-2013 destabilizes the folded state by ∼2 kcal/mol. ΔG_{fold} is within error of the experimental

range for C22* and is just outside this range for the Drude-2013 model, but is far too negative for C36. Therefore, the C22* and Drude-2013 models are roughly equivalent in describing the folding behavior of the GB1 β-hairpin, with the C36 model performing the worst of the three force fields.

Table 5.1: ΔG_{fold} *for GB1.* Calculated from 1D PMFs in units of kcal/mol.

Coordinate	C36	C22*	Drude
N_{hb}	-7.22 ± 2.15	-1.73 ± 2.16	$+2.77 \pm 1.74$
R_{G}	-7.85 ± 3.07	-2.85 ± 1.26	$+1.65 \pm 3.11$

5.3.2 Side-Chain Hydration Energies

To examine the differences in GB1 stability between the three force fields, we calculated hydration energies of amino-acid side-chain analogs using the C36 and Drude-2013 models (Table 5.2). With the exception of Asp, Glu, Arg, and Trp side-chain partial charges, C22* non-bonded parameters are equivalent to C36 [144, 167]. In the analogs used, the side chain is terminated with a hydrogen atom at the position where the α-carbon would be bonded. Molecules constructed to mimic the side chain of an amino acid do not necessarily serve as an appropriate measure of the hydration of a side chain attached to a protein [168], but they do serve to test if the non-bonded parameters for the chemical moieties present in the side chains are appropriate for describing their interactions with water [169, 170, 171].

In general, hydration energies predicted by both the C36 and Drude-2013 models are in good agreement with the experimental hydration energies. The C36 and Drude-2013 models both show a significant improvement in Trp side chain hydration over C22 (equivalent to C22*), which underestimates the free energy by 3 kcal/mol [158], whereas the C36 and Drude-2013 models only underestimate it by 0.5 and 0.8 kcal/mol, respectively. The Drude-2013 model also shows a systematic improvement for sulfur-containing side chains, Cys and Met, over the C36 model. The Drude-2013 model correctly predicts these to be sparingly soluble, while the C36 model underestimates their solubility by 1 and 2 kcal/mol, re-

Table 5.2: *Hydration energies of amino acid side chain models.* Calculated using the C36 non-polarizable model and the Drude-2013 polarizable model. For WCA decomposition and errors, see Tables B.1 and B.2. The experimental hydration energies are from Refs. [172] and [173].

molecule	residue	ΔG_{C36}	$\Delta G_{\mathrm{Drude}}$	$\Delta G_{\mathrm{exptl}}$
n-butane	Ile	2.60	2.68	2.08
isobutane	Leu	2.56	2.50	2.28
methane	Ala	2.38	2.24	2.00
propane	Val	2.49	2.65	1.96
acetamide	Asn	−7.56	−9.69	−9.72
p-cresol	Tyr	−4.92	−5.52	−6.13
ethanol	Thr	−4.78	−3.74	−4.90
methanethiol	Cys	−0.27	−1.04	−1.24
methanol	Ser	−4.94	−3.43	−5.08
methylethylsulfide	Met	0.59	−1.12	−1.49
3-methylindole	Trp	−5.40	−5.11	−5.91
methylimidazole	His	−10.11	−11.67	−10.25
propionamide	Gln	−7.59	−8.06	−9.42
toluene	Phe	−0.30	−0.48	−0.76
acetate	Asp	−82.35	−84.92	−80.65
n-butylammonium	Lys	−66.00	−61.90	−69.24
n-propylguanidinium	Arg	−60.21	−57.00	–
propionate	Glu	−82.75	−80.76	−79.12
n-methylacetamide	NMA	−7.65	−8.50	−10.10

spectively. Conversely, the hydration energies of non-aromatic hydroxyl containing groups (i.e., Ser and Thr) are underestimated by the Drude-2013 model, while the C36 model predicts them accurately. This difference appears to stem from the electrostatic component of the hydration energy: in the Drude-2013 model they are roughly 1 kcal/mol smaller than those of the C36 model (see Tables B.1 and B.2). This is consistent with the O-H bond in the Drude-2013 model being less polar ($q_H = 0.36$e in the Drude-2013 model, while $q_H = 0.43$e in the C36 model). Polar and non-polar residues alternate along the two strands of the GB1 β-hairpin, a common motif in anti-parallel β-strands [174], which situates the polar and non-polar residues to point in the opposite direction, leading to distinct polar and non-polar faces of the β-sheet. Given that four of the six hydrogen-bonding residues in GB1 are threonines, reduced solvation of the threonine side chains in the Drude-2013 model could have a compounding effect, reducing the hydration free energy of the polar face of the β-hairpin.

The C36 model underestimates the hydration energies of amides (i.e., NMA, Gln, and Asn) by roughly 2 kcal/mol. One justification for this underestimation is that the force field compensates for the lack of induced polarization in the model. When a protein folds into secondary structures, hydrogen bonding between backbone amides results in polarization of the C=O and N-H bonds, which gives a complementary effect that favors the propagation of secondary structures [143]. Non-polarizable force fields like C36 are incapable of describing this effect rigorously, but by defining the water–amide Lennard-Jones parameters such that the interactions are weakened, the model has the correct propensity for forming secondary structures.

In contrast to the C36 model, the Drude-2013 model hydration energies of the amide-containing molecules (i.e., NMA, Gln, Asn) are in better agreement with the experimental values. This suggests a possible explanation for the Drude-2013 model predicting a (slightly) unstable β-hairpin, as opposed to a stable β-hairpin for C36. The amide hydration energies of the Drude-2013 model are more negative than the C36 model by roughly

1 kcal/mol. If the stabilization that is achieved by the backbone–backbone hydrogen bonds of the β-hairpin does not compensate for the increased hydration energy of the amides, the unfolded state will be favored.

5.3.3 Adjusting Drude Polarizabilities

We observed key differences in hydration free energies between the Drude-2013 and C36 models that lead to distinct free energy profiles for GB1 β-hairpin folding. The Drude-2013 model improves the hydration free energies over C36 for some chemical motifs, namely amides and sulfur-containing compounds, while it underestimates the free energies for other motifs, such as hydroxyl groups. Although the NMA hydration free energy is improved in the Drude-2013 model, it is still underestimated by $\sim$1.5 kcal/mol relative to the experimental value. Enhancing the atomic polarizability of backbone amide N and side chain hydroxyl O atoms should improve the hydration of peptide backbone and Thr and Ser side chains, respectively. We recalculated the electrostatic component of the hydration free energies for NMA, Thr, and Ser for atomic polarizabilities of amide N and hydroxyl O atoms scaled by up to 2$\times$ (see Fig. 5.3).

For NMA, experimental hydration free energies are most accurately reproduced when the polarizability of the amide N atom is increased by 60%. For Ser and Thr side chains, increasing the polarizability of the hydroxyl O atom by 50% and 30%, respectively, were the most accurate in reproducing experimental hydration free energies. For all of NMA, Ser, and Thr, increasing their respective atomic polarizabilities by 40% reproduces experimental hydration free energies within $\sim$0.5 kcal/mol. Adjusting the atomic charge of the amide H atom to its C36 value also reproduced NMA hydration free energies fairly well, underestimating the free energy by less than 0.5 kcal/mol; adjusting the amide N atomic charge to its C36 value, however, overestimates the hydration free energy by more than 1 kcal/mol (see Supplementary Material). Overall, side chain hydroxyl free energies are more sensitive to changes in polarizability and dipole moment than amides. This might

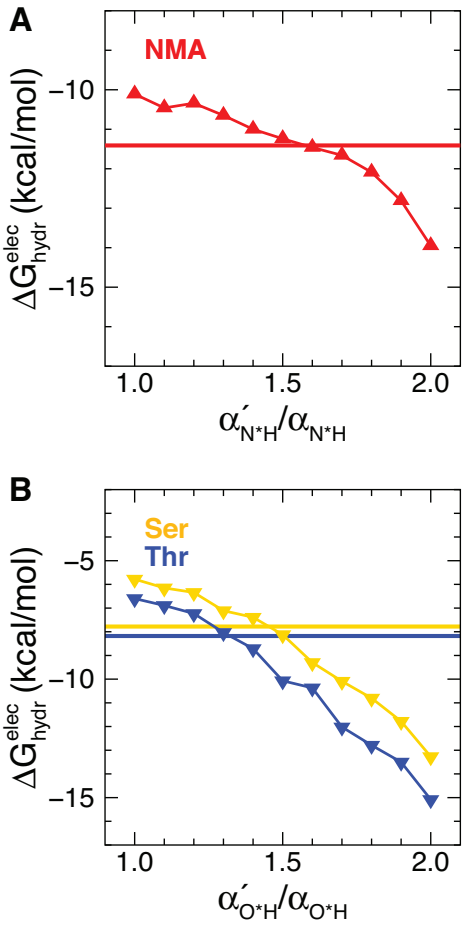


Figure 5.3: *Electrostatic component of the hydration free energies.* (A) For NMA, the polarizability of the amide N ($\alpha'_{\mathrm{N}*\mathrm{H}}$) was adjusted from 1-2$\times$ of its Drude-2013 value ($\alpha_{\mathrm{N}*\mathrm{H}}$) (red, upward-facing triangles). The solid red line shows the target electrostatic component that would reproduce the experimental hydration free energy for NMA. (B) For Thr (blue) and Ser (gold), the polarizability of the hydroxyl O ($\alpha'_{\mathrm{O}*\mathrm{H}}$) was adjusted from 1-2$\times$ of its Drude-2013 value ($\alpha_{\mathrm{O}*\mathrm{H}}$). The solid lines show the target electrostatic component that would reproduce the experimental hydration free energies for Thr (blue) and Ser (orange), respectively.

be expected since, although N atoms are more polarizable than O atoms, O-H groups have larger intrinsic dipole moments than N-H groups. In addition, C=O groups have a large effect on the solubility of amides because they have larger dipole moments, both intrinsically and when hydrogen bonded to water, than the N-H groups (see Fig. B.6).

We recalculated our folding PMFs for GB1 using these new polarizability parameters (see Methods and Fig. 5.4). We extended the range of polarizabilities tested for the N and O atoms down to 0 in order to better understand the role of polarizability in the folding behavior. As might be expected, the folded state becomes more favorable as the backbone amide N atom becomes more polarizable, with ΔG_{fold} decreasing roughly monotonically with increasing polarizability, becoming negative when the polarizability is increased by 70% ($1.7\times$) compared to the original Drude-2013 parameter. For increases of 30-60% ($1.3 - 1.6\times$), ΔG_{fold} decreases by $\sim$1.4-1.9 kcal/mol over the original polarizability, with 30% showing the largest drop, bringing ΔG_{fold} closer to the experimental values of 0.0-0.5 kcal/mol [122, 130]. A similar shift is seen when adjusting the atomic partial charges of the amide N and H atoms to their C36 values (see Supplementary Material). Because using reduced data sets can introduce a sampling bias to the free energy (see Methods), we approximate this bias by subtracting the full-data free energy from the reduced-data free energy: $\Delta G_{\text{fold}}^{\text{bias}} = \Delta G_{\text{fold}}^{\text{reduced}} - \Delta G_{\text{fold}}^{\text{full}}$ = $+1.4$ kcal/mol. If we adjust for this bias, increasing the polarizability 30-60% should produce ΔG_{fold} values only $\sim$0.5-1.0 kcal/mol greater than experimental values, which is within the error of the PMF calculations (see Table 5.1). However, the change in parameters could be affected differently by the sampling bias, meaning that full REUS calculations would be needed to verify the improvement in the free energies. From our perturbative approach, either a 30-60% increase in the amide N polarizability or a reversion of the backbone amide H atomic charge to its C36 value appears to be a viable means of improving both the backbone solubility and β-hairpin folding properties in the Drude-2013 model. Alternatively, decreasing the Lennard-Jones minimum between the backbone amide N and carbonyl O atoms by 7% also reproduces

experimental folding free energies, but without the additional improvement in solubility (see Supplementary Material).

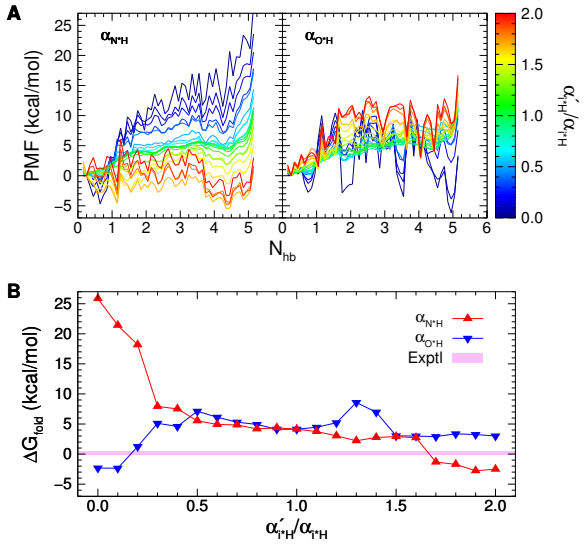


Figure 5.4: *Free energies with adjusted Drude-2013 polarizabilities and charges.* (A) 1D PMFs of GB1 along the N_{hb} coordinate for adjusted (left) N-H and (right) O-H parameters using 10-ps snapshots from last 5 ns/window of REUS simulations (see Methods). For original Drude-2013 parameters ($\alpha'/\alpha = 1$), the PMF was also recalculated using this reduced data set. (Solid lines) 1D PMFs for adjusted N or O atomic polarizabilities, colored by ratio between new (α') and old (α) polarizabilities. (B) Folding free energies calculated from 1D PMFs in (A) (see Methods). The red and blue triangles are for modified amide N and hydroxyl O polarizabilities, respectively. The pink band gives the experimental folding free energies of GB1 from Refs. [122] and [130].

Unlike the backbone amide N polarizabilities, altering side chain hydroxyl O polarizabilities does not monotonically change the folding free of GB1. GB1 contains five Thr residues, four of which are involved in native β-hairpin hydrogen bonding, with the fifth residing in the turn region. Increasing the polarizability of Thr side chain hydroxyl O atoms by 30%, which most accurately reproduces experimental hydration free energies (see Fig. 5.3B), results in the largest, positive folding free energy for all tested polarizabilities, shifting it to the unfolded state by $+4.4$ kcal/mol over the original Drude-2013 parameters. At larger polarizabilities the folding free energy stabilizes to around 1 kcal/mol lower than the original polarizability. On the other hand, reducing the polarizability of the hydroxyl O atom close to 0, i.e., making the Thr side chains more hydrophobic, stabilizes the folded state to the same extent as greatly increasing the backbone amide N polarizability.

The increase in folding free energy around $1.3\times$ the original O polarizability appears to be driven by an increase in free energy of the intermediate region between N_{hb}=2-4. For higher polarizabilities this is further accompanied by a corresponding decrease in the free energy near the folded state. A more modest increase of 20% in the polarizability shifts the folding free energy up by 1 kcal/mol over the original parameters while underestimating the Thr side chain hydration free energy by only 1 kcal/mol. This appears to be a reasonable trade off, especially if combined with the aforementioned adjustments to the backbone amide polarizability or atomic partial charges.

We tested the proposed changes to the Drude-2013 parameters for an α-helical peptide, Ala_{10}, a decamer of alanines (see Supplementary Material for more details). Ala_{10} folding free energy landscapes for the three CHARMM force fields tested here follow roughly the same patterns observed for GB1 folding, namely that the Drude-2013 model favors the unfolded, extended state more so than the additive C36 and C22* models, with $\Delta G_{\mathrm{fold}} > 0$ for all three models (see Fig. B.8 and Table B.3). For the α-helical peptide, $(\mathrm{AAQAA})_3$, the Drude-2013 model correctly predicted a positive folding free energy (only $\sim$25% folded), although this was actually slightly higher than for C36 ($\sim$20% folded) [143]. For Ala_{10} we can roughly compare the folding free energies to those of Ala_5. Graf et al. showed that backbone NMR coupling parameters were roughly constant for Ala_3 up to Ala_7, and fitting MD simulations to the NMR results showed very little change in the distribution of structures between Ala_5 and Ala_7, with no α-helical content in either [115]. Best et al. repeated this analysis for Ala_5 with a larger sample of MD simulations using multiple force fields and found it to contain $\sim$2-11% helical content ($\Delta G_{\mathrm{fold}} \sim 1.2 - 2.3$ kcal/mol). For Ala_{10} the Drude-2013 model sits right at the upper bound of this range, while C36 and C22* sit just below this range, although they are all within one standard deviation of it. Only small increases in amide N polarizabilities of 10-20% marginally decrease ΔG_{fold}, while all others cause ΔG_{fold} to increase. We see similar results for the Lennard-Jones parameters, where a decrease in the N- -O LJ minima of up to 4% yields small decreases

in ΔG_{fold}, with all larger deviations increasing ΔG_{fold}. C36 backbone amide charges also leave the folding free energy outside of the Ala_5 range. This suggests, along with good experimental agreement for $(\text{AAQAA})_3$, that the Drude-2013 parameters are optimized for α-helices, but may be less so for β-sheets. A modest increase in backbone amide N polarizabilities of 30% could be sufficient for improving both the backbone hydration and β-sheet folding free energies while not appreciably affecting α-helical folding free energies.

5.3.4 Other Avenues for Improvement of the Drude Force Field

Separate from the issues with folding GB1, the amino-acid-analog hydration energies suggest other areas where the parameters for the Drude-2013 force field could be improved. As noted above, the hydration energies of the hydroxyl-containing amino acids are underestimated by the Drude-2013 model. The parameters for the charged side chains could also be improved; the hydration energies of Glu and Asp are overestimated while the hydration energy of Lys is underestimated. Although there is no experimental value to compare to, the Drude-2013 model hydration energy for Arg is 4 kcal/mol lower than is predicted by the C36 model.

The Drude-2013 model hydration energy of Asn is more negative than that of Gln (-9.7 kcal/mol vs. -8.1 kcal/mol), although the experimental hydration energies are essentially equal (-9.7 kcal/mol and -9.4 kcal/mol, respectively). This difference results from the use of a smaller Lennard-Jones radius for the C_β of Asn (atom type CD32C, σ=3.22 Å) in comparison to Gln (atom type CD32A, σ=3.74 Å); when the hydration energy of Gln is calculated using the CD32C atom type for C_β, it becomes -11.2 kcal/mol. The σ parameter for the Asn CD32C atom type is actually anomalously small for an aliphatic carbon, which results in a smaller repulsive component of the hydration energy. This suggests that the Drude-2013 model hydration energy of Asn only agrees with the experimental value because of the small radius of C_β, but it does not accurately represent the strength of the

water–amide interactions.

5.3.5 A Non-native Salt Bridge Overstabilizes the GB1 β-Hairpin in the C36 Model

We also examined interactions in the β-turn region of GB1 and how they correlate to folding. Salt bridges that stabilize the turn have been shown to increase the stability of the β-hairpin for another model β-sheet peptide, the PIN1 WW domain [175]. For GB1 however, the Asp47–Lys50 salt bridge in the turn, although appearing in some crystal structures, does not consistently form in solution [176]. For C22* and Drude-2013, the salt bridge does not form, while for C36, it forms before hydrogen bonding begins and is maintained throughout the folding process (see Fig. 5.5). This salt bridge appears to be the source of the overstabilized β-hairpin in the C36 model. In the Drude-2013 model, Lys50 dipole moments are significantly reduced compared to C36 (a decrease of $\sim$3-4 D), while other, non-glutamate residues remain roughly the same between the three force fields (see Fig. B.5). Interestingly, Asp47 dipole moments are also significantly reduced in C22* compared to C36, most likely due to decreased partial charges of the aspartate side-chain carboxylate group in C22* (Ref. [144]) as a similar decrease in dipole is also observed for Asp46 (see Fig. B.5). The decrease in dipole moments for Asp47 and Lys50 in the C22* and Drude-2013 models, respectively, from their C36 levels could explain why the salt bridge is only observed in the C36 model.

5.3.6 Tryptophan Fluorescence Does Not Capture the Full β-Hairpin Folding Pathway

Lastly, we wanted to address experimental measures of protein folding and how this compares to folding observed in MD simulations. Muñoz et al. measured GB1 folding using fluorescence measurements of the buried Trp43 side chain [122]. Solvent accessible surface area (SASA) correlates well with tryptophan fluorescence [177], so we calculated the SASA of Trp43 for our REUS simulations. As one might expect, Trp43 SASA correlates well with the R_{G} coordinate but poorly with the N_{hb} coordinate (see Fig. 5.6). If

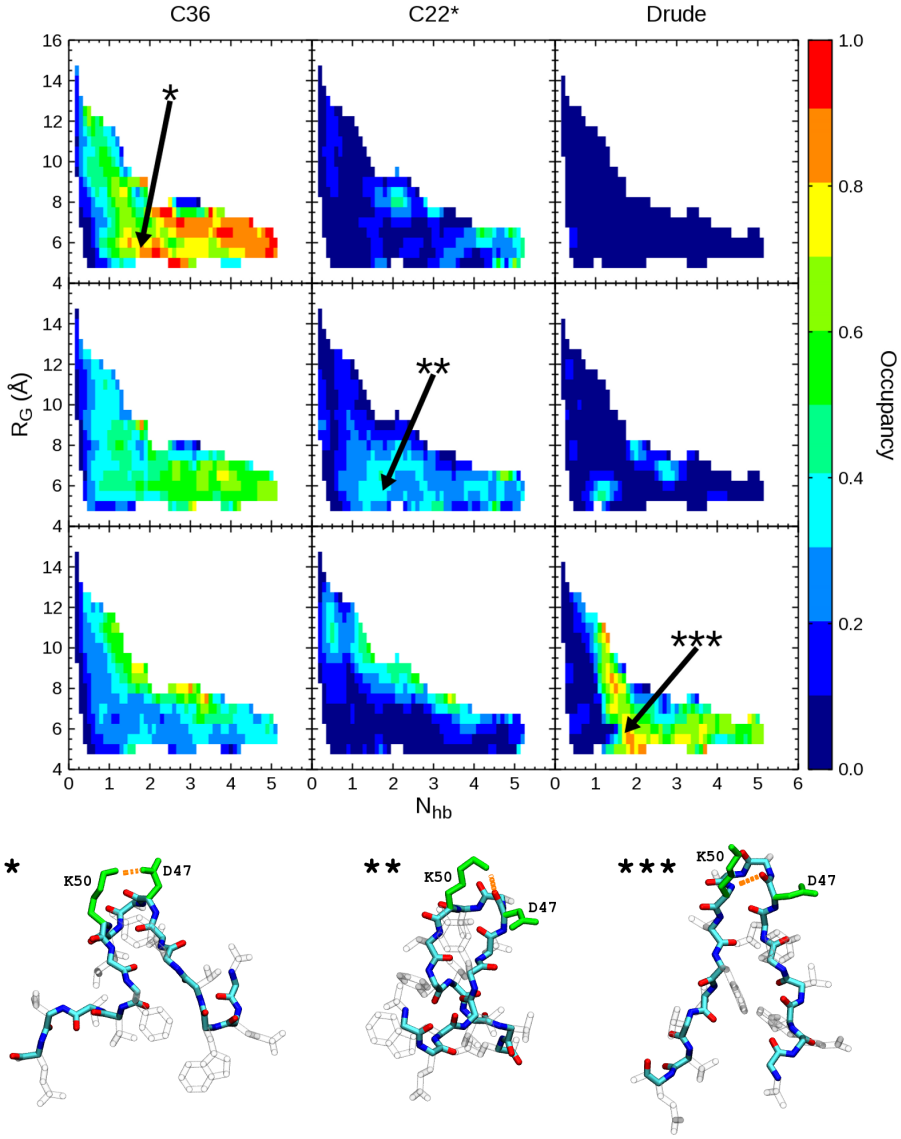


Figure 5.5: *Asp47–Lys50 interactions in the turn region of GB1.* Occupation of (top) Asp47–Lys50 salt bridge, (center) Asp47 backbone–Lys50 side chain hydrogen bond, and (bottom) Asp47 backbone–Lys50 backbone hydrogen bond for (left) C36, (middle) C22*, and (right) Drude-2013 REUS trajectories. Salt Bridges and hydrogen bonds were considered occupied when the distance between their respective atoms was <3.4 Å. Representative structures are shown in licorice representation for (*) C36, (**) C22*, and (***) Drude-2013, with their coordinates indicated in the plots. Backbone atoms are colored by atom name. Asp47 and Lys50 side chains are shown in green, while all other side chains are transparent. Hydrogen bonds and salt bridges are shown as orange, dashed lines.

we integrate the 2D PMFs over the fraction of exposed Trp43 surface area, we see only a single well for all three models (see Fig. B.11). This follows what was observed for the R_{G} coordinate, where only the original Drude-2013 model showed two-state behavior (see Fig. B.1). Essentially, since formation of the hydrogen bonds follows the hydrophobic collapse, the exposed tryptophan surface area only tracks the hydrophobic collapse portion of the folding process, giving no indication of hydrogen bond formation. This agrees well with experiments that have found that folding rates determined by tryptophan fluorescence measurements are often faster than the true full folding rate [178]. Thus, tryptophan fluorescence alone is not a sufficient reaction coordinate for protein folding and, instead, requires at least one supplemental coordinate to capture the complete folding process.

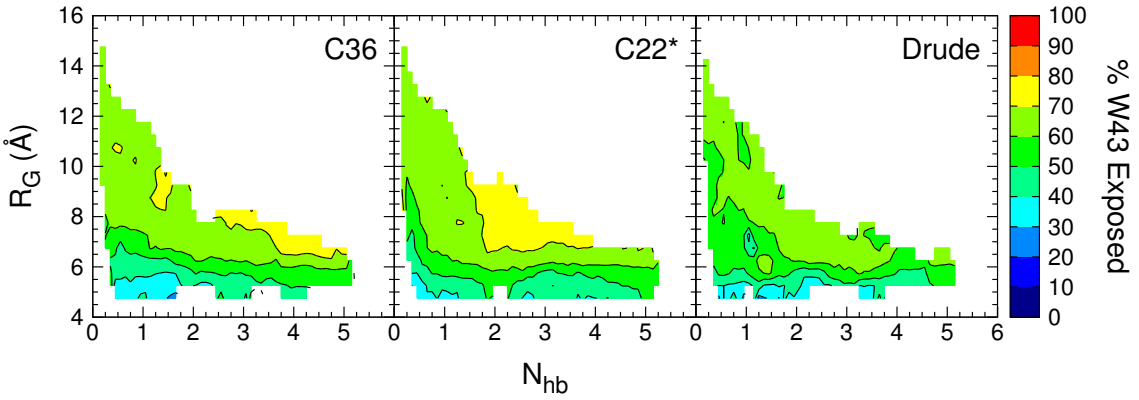


Figure 5.6: *Fraction of exposed surface area for Trp43 side chain.* Fraction of exposed surface area of Trp43 side chain was calculated by dividing the solvent-exposed surface area by the total surface area. Contours at 10% intervals are shown in black.

5.4 Conclusions

We calculated folding free energy landscapes of the GB1 β-hairpin using CHARMM non-polarizable force fields and the CHARMM Drude-2013 polarizable model and found that the Drude-2013 model destabilizes the β-hairpin, resulting in $\Delta G_{\mathrm{fold}} \sim +2$ kcal/mol, versus -7 and -2 kcal/mol for C36 and C22*, respectively. Hydration free energies show improved solvation of the peptide backbone in the Drude-2013 model over the non-polarizable C36, suggesting that peptide–peptide hydrogen bonding is too weak to compensate for a

more solvated backbone. By increasing the backbone amide N atomic polarizabilities by 30%, we were able to further improve the backbone solubility as well as shift the folding free energy closer to the experimental value of $\sim$+0.5 kcal/mol [122, 130].

Improved backbone hydration energies in the Drude-2013 model may be the result of numerous occurrences. First, C36 amide hydration energies are underestimated by $\sim$2-3 kcal/mol (see Table 5.2). This presumably compensates for the lack of N-H and C=O polarization during hydrogen bonding. However, although this works well for α-helices [34], it results in overstabilized β-sheets. Altered dihedral parameters and the removal of the helix-overstabilizing CMAP potential allows C22* to avoid such a strong preference for the folded state [118, 144]. For the Drude-2013 model, altered partial charges combined with dynamic atomic polarization strengthen backbone–water hydrogen bonding. O partial charges are more negative in the SWM4-DNP water model (-0.834e for TIP3P compared to -1.115e for SWM4-NDP), and they have a deeper Lennard-Jones potential well (-0.15 kcal/mol for TIP3P compared to -0.21 kcal/mol for SWM4-NDP), both of which are typical for 4-point water models [179, 180, 181]. Additionally, the net charge for backbone N-H groups increases from -0.16e in C36 to -0.11e in Drude-2013, while the net charge for backbone C=O groups decreases from 0e in C36 to -0.04e in Drude-2013. The reduced charge for N-H groups is due to a less negatively charged N atom and a less positively charged H atom, while the net negative charge for C=O groups is a result of a less positively charged C atom and a more negatively charged O atom, with the negative charge on the O atom split between two lone pair charges. N-H and C=O dipole moments are also typically higher in the Drude-2013 model than in the C36 or C22* models. For native hydrogen-bonding residues, N-H dipole moments in the Drude-2013 model are enhanced by polarization of the N atoms and alignment of the dipoles during folding. Conversely, C=O dipole moments are enhanced primarily by alignment of the atomic dipoles, as the C and O atoms are less polarizable than the N atoms. C and O atoms do polarize to a larger extent when involved in water–peptide hydrogen bonding, however. As we also observed

with hydroxyl groups in polar side chains, the reduced partial charges on the backbone N and H atoms in the Drude-2013 model may be decreasing its affinity for hydrogen bonding. This would suggest that improvements in amide hydration free energies in the Drude-2013 model are due primarily to stronger carbonyl–water interactions. Increasing the N-H partial charges and enhancing N polarizabilities were both successful in improving the solubility of the backbone, with the latter also stabilizing the GB1 β-hairpin.

Adjusting parameters in any model should be performed with care. Atomic polarizabilities in the Drude-2013 model, for example, have already been scaled from their gas-phase values by 0.85 to reproduce experimental values for the dielectric constants of pure liquids of pyridine and pyrrole [33, 182]. Our proposed increase of 30% for Drude-2013 backbone amide N polarizabilities would increase them by 10.5% over their gas-phase values. While the dielectric constant for pyridine, which does not have an N-H group, was unchanged between gas-phase scaling factors of 1.0 and 0.85, the dielectric constant for pyrrole, which does have an N-H group and should be more representative of backbone amide values, was overestimated by $\sim$20% with a scaling factor of 1.0 [182], suggesting that our polarizability changes would destroy the improvement in bulk properties with the 0.85 scaling factor. However, while increasing the atomic polarizabilities above the 0.85 scaling factor clearly has adverse effects on the dielectric constant of pyrrole, changes in the dipole and quadrupole moments were much less significant, differing by only 2-4% between 0.85 and 1.0, with the higher scaling factor producing quadrupole moments slightly closer to quantum mechanical calculations [182]. Since we only increase the polarizability of protein backbone amide N atoms while keeping all other protein atoms at their 0.85 scaled values, the overestimation of the pyrrole dielectric constant should be limited. In addition for unfolded states of proteins, the bulk properties of water, which we do not change, should be more important than the bulk properties of the protein backbone. Therefore, overestimation of the pyrrole dielectric constant becomes less important in this context. For the folded states of larger proteins, where the dielectric constant near the center of the protein

is much smaller than for bulk water and much more dependent on the electrical properties of the amino acids [183], changes in the dielectric constant of the backbone will be more significant than for the small β-hairpin of GB1. Studies on larger proteins will be necessary to determine the effects of polarization on buried residues. Lastly, a major advantage of the Drude-2013 model is the avoidance of the overpolarization present in many fixed charge force fields, which often leads to overstabilization of secondary structures. It accomplishes this by introducing dynamic atomic polarization while simultaneously reducing some of the intrinsic polarization. We observed that overpolarization in the Drude-2013 model can also lead to overstabilization of secondary structure, although this appears to be system dependent as opposite effects were observed the GB1 β-hairpin and the Ala_{10} α-helix.

The C36 model greatly overstabilizes the GB1 β-hairpin due to a non-native salt bridge that occurs in the turn region. Elimination of this salt bridge allows C22* to only slightly overstabilize the β-hairpin. Weaker interactions with the TIP3P water model compared to the SWM4-NDP water model in the Drude-2013 systems along with stronger peptide–peptide hydrogen bonding, as well as a more hydrophobic tryptophan side chain, appear to be the stabilizing factors for C22*. Ultimately, the C22* and Drude-2013 models produce roughly equivalent folded fractions that are within or nearly within the experimental range of 30-50% [122, 130]. In the recently released C36m force field, the CMAP potential in the C36 force field was refined to more accurately reproduce protein unfolded states observed in experiments [184]. However, rescaling our C36 REUS trajectories with the C36m force field parameters according to Eq. 5.4 produced no discernible difference in the folding free energy landscape of GB1 (see Fig. B.12), in agreement with the finding that C36 and C36m sample similar conformational ensembles for GB1 [184]. The C36m force field also includes a pair-specific Lennard-Jones parameter to weaken salt bridge interactions between side chain guanidinium N (Arg) and carboxylate O (Asp/Glu) atoms to correct for an overestimation of the equilibrium assocation constant of a guanidinium-acetate solution and an underestimation of its osmotic pressure [184]. This was proposed as

a simpler, less perturbative method to that employed by Piana et al. in the C22* force field, which reduced the partial charges of the guanidinium and carboxylate groups to achieve the same effect [144]. The additional carboxylate–ammonium-specific Lennard-Jones parameter could improve the folding properties of GB1 in the C36m force field by preventing formation of the non-native Asp47–Lys50 salt bridge.

Lastly, we observe that tryptophan fluorescence, an oft-used method for studying protein folding, does not properly capture the full folding pathway of the GB1 β-hairpin. For all three CHARMM models studied here, hydrophobic collapse precedes hydrogen-bond formation, and the model of hydrophobic contacts proposed by Muñoz et al. does not hold for our simulations [132]. GB1 folds via a "zippering" mechanism, with hydrogen bonds in the turn region forming first and hydrogen bonds near the termini forming last. In this mechanism, tryptophan fluorescence would underestimate the true folding rate, since the energetic barrier for folding is hydrogen bond formation, not hydrophobic collapse. Another method that monitors non-native contacts is an alternative approach that has been proven to correctly measure folding rates [178]. Additionally, one could combine tryptophan fluorescence with another technique that also captures the zippering mechanism observed in our simulations, such as FRET to measure contacts near the termini.

CHAPTER 6

THE VECTORIAL PATHWAY SIGNIFICANTLY ENHANCES THE *IN VIVO* FOLDING RATE OF β-HELICAL PASSENGER DOMAINS OF AUTOTRANSPORTERS IN PATHOGENIC GRAM-NEGATIVE BACTERIA

6.1 Introduction

Autotransporters are a class of proteins in Gram-negative bacteria that are typically involved in bacterial virulence. They are so-named due to the prevailing theory that they contain all the functional elements needed for secretion of their virulence factors, although this may not be entirely true. Autotransporters consist of two main domains, a C-terminal 12-stranded β-barrel ("β") domain [185, 186, 187], which resides in the outer membrane (OM), and an N-terminal β-helical ("passenger") domain [6], which gets threaded through the β-domain via the type Va secretion mechanism and contains the virulence factor itself [188]. How this secretion happens is highly debated. It is known that passenger domains are secreted unidirectionally, C- to N-terminus, through the β-domain [5, 189], and it has been proposed with some experimental evidence that secretion of the passenger domain occurs via a vectorial folding pathway, in which small β-structures form at the C-terminus after secretion that prevent backsliding of the peptide strand through the β-domain [5, 6, 190, 191]. Other experiments have challenged this theory, showing both that folding of the passenger domain is not necessary for secretion [192] and that the Bam complex orchestrates membrane insertion of the β-domain and secretion of the passenger domain in a concerted process [189, 193, 194]. Given the absence of ATP or other traditional chemical energy sources at the outer membrane, the energetics of passenger domain folding was attributed as the energy source. Energetic analysis, however, suggests that this folding energy is but a small fraction of the energy required for secretion [192].

The autotransporter pertactin from *Bordetella pertussis*, the causitive agent of whooping cough, is one oft-studied model of passenger domain secretion. Patricia Clark and coworkers observed that *in vitro* folding of the pertactin passenger domain occurs over a time span of several hours, while taking several days to reach equilibrium [6]. This timescale for folding is much longer than what is expected *in vivo*, given the typical doubling time of pathogenic bacteria is $\sim$30-60 min. Indeed, experiments where OM secretion was stalled showed pertactin to be unfolded while still in the periplasm after $>$90 min, yet it immediately folded once secretion was resumed [195]. Furthermore, the N- and C-terminus of the helix fold at the same rate *in vitro* – the same rate as the entire passenger domain – indicating a concerted folding process [7]. An off-pathway, partly folded intermediate state consisting of the first six C-terminal rungs of the β-helix, called the C-terminal core, was observed to be stable on its own and was initially thought be a scaffold for further folding *in vivo* [6, 196]. However, its folding rate turned out to be the same as the full passenger domain, so its role as the scaffold was abandoned [7]. An extracellular loop on the β-domain may instead serve this role [197], while a hydrophobic cavity in the extracellular pore of the β-domain also promotes folding of the passenger domain [198]. Therefore, the protein folds along a very different pathway during secretion than it does *in vitro*, and the cellular environment dictates which pathway the protein follows.

Kinetic models predict passenger domain folding to be the driving force for secretion efficiency [195]. Interestingly, destabilizing the N-terminus of the passenger domain increases efficiency, while destabilizing the C-terminus has the opposite effect [190, 191, 199]. Additionally, in the absence of the C-terminus, the N-terminus remains unfolded [200]. These two results imply that a free energy gradient along the passenger domain β-helix appears to be important for efficient secretion. This free energy gradient could serve three purposes: (i) to prevent folding during IM secretion into the periplasm, which proceeds in the opposite direction of OM secretion, (ii) to slow the formation of the native state while in the periplasm, and (iii) to facilitate the faster, vectorial pathway

during OM secretion into the extracellular space. In this study, we use molecular dynamics simulations and enhanced sampling methods to compare the energetics of the *in vitro* and *in vivo* folding pathways of the pertactin passenger domain in order to assess the efficacy of the vectorial pathway as the energy source for efficient secretion.

6.2 Methods

6.2.1 System construction

The pertactin β-helical passenger domain structure was taken from Ref. [201] (PDB code: 1DAB). We built systems to measure the potential of mean force (PMF) for the folding of individual bands of the β-helix both in isolation and along the vectorial folding pathway. Here a "band" refers to two successive "rungs" of the β-helix; e.g, band1C refers to the first two rungs, or β-strands, of the β-helix which form a single, parallel β-sheet. CHARMM22* [144] parameters were chosen to describe the protein, which we have previously shown to accurately describe β-sheet folding [202].

To measure the folding PMFs for isolated bands, we isolated each band from the rest of the structure. Table 6.1 details which residues were used for each band system. Each peptide was then solvated in a 150 Å $\times$ 70 Å $\times$ 70 Å box with $\sim$23,000 TIP3P water molecules [92], and then neutralized with 0.15 M NaCl, for a total system size of $\sim$70,000 atoms. Additional restraints were added to align the end-to-end vector the peptide along the x-axis.

For vectorial folding PMFs of each band, we constructed systems which also included all C-terminal residues to that band, stopping at residue 280 for bands 1N, 2N, and 3N and residue 539 for bands 1C, 2C, 3C, and the kinked band. Each system was solvated in a water box with lengths chosen to accommodate the peptide with a fully unfolded and extended N-terminal rung. This resulted in box lengths of $\sim$65-80 Å for the y- and z-dimensions and 117-193 Å for the x-dimension. Each system was also neutralized with 0.15 M NaCl. Total system sizes ranged from $\sim$53,000-106,000 atoms. The heavy backbone atoms of all β-

Table 6.1: *Isolated bands of the pertactin β-helical passenger domain.*

Band	Residues	Native HBonds	Initial # of windows	Final # of windows	SLUS iterations
1N	1-38	7	187	414	17
2N	16-64	16	234	769	14
3N	45-96	17	269	698	14
Kinked	262-313	18	267	816	9
3C	396-428	11	226	560	10
2C	411-452	12	293	542	9
1C	438-481	16	252	589	9

sheet residues not in the last, N-terminal rung were restrained to their native conformations.

6.2.2 Molecular dynamics simulations

Molecular dynamics simulations were carried out using NAMD 2.10-12 [**Phillips2005**]. The temperature was fixed at 310 K using Langevin dynamics; the pressure was kept constant at 1 atm using the Langevin piston method [96]. The equations of motion were integrated using the RESPA multiple time-step algorithm with a time step of 2 fs used for all bonded interactions, 2 fs for short-range non-bonded interactions, and 4 fs for long-range electrostatic interactions. Long-range electrostatic interactions were calculated using the particle-mesh Ewald method with a real-space cutoff of 12 Å [97]. Short-range non-bonded Lennard-Jones interactions were cutoff at 12 Å with a potential switching function beginning at 10 Å bringing the potential energy smoothly to zero at the cut off distance. Bonds involving hydrogen atoms were constrained to their equilibrium length, employing the SETTLE algorithm [146] for water molecules and the SHAKE algorithm for all others [147].

6.2.3 Self-learning adaptive umbrella sampling procedure

Two-dimensional folding free energy landscapes of the β-helical bands of pertactin were calculated using umbrella sampling (US). Our two reaction coordinates are similar to those

previously used for the β-hairpin peptide, GB1 [202], namely the number of native hydrogen bonds along the β-sheets, N_{hb}, and the radius of gyration for all internal hydrophobic side chains, R_{G}. The N_{hb} coordinate is the sum of individual hydrogen bond collective variables, defined by Eq. 5.1. This hbf function is a smooth function where if two atoms are well within the cutoff (3.3 Å), then hbf ~ 1, and ~ 0 when they are well outside the cutoff. Windows were spaced by 0.25 along N_{hb}, with a harmonic force constant of 25.0 kcal/mol, and by 0.5 Å along R_{G}, with a harmonic force constant of 6.25 kcal/mol·Å^2. All US simulations were run at 310 K. PMFs were generated via the weighted histogram analysis method (WHAM) [45, 155].

To initiate the self-learning adaptive umbrella sampling (SLUS) procedure [8] for the isolated bands, we first ran a 10-ns simulation at 310 K followed by another 10-ns simulation at 500 K. From these simulations we generated starting states for a 2D umbrella sampling simulation. We used a 5 kcal/mol free energy cutoff for our SLUS procedure, with 2 ns/window simulations for each iteration. We used replica exchange umbrella sampling (REUS) [51] for the initial set of windows (iteration 0), and normal US for all subsequent iterations. Upon completion of the SLUS procedure, we then ran 10 ns/window of REUS on all generated windows. Table 6.1 tabulates the number of SLUS iterations and the number of generated windows for each band.

For the vectorial folding PMFs, we used $\sim$50-150-ns metadynamics simulations [41] for each system to generate the starting windows. Since the metadynamics simulations cover most of the reaction coordinate space already, the initial round of REUS simulations were run for 7 ns/window, with the last 5 ns/window used for calculating the PMFs.

6.3 Results

6.3.1 Instability of the isolated bands of the pertactin β-helical passenger domain revealed through potentials of mean force

We used self-learning adaptive umbrella sampling (SLUS; see Methods) [8] to explore the folding free energy landscapes of individual isolated bands of the pertactin β-helical passenger domain (see Methods for definition of the bands). We examined the first three bands at the C- and N-termini, as well as a central kinked band that consisted of two full rungs with one incomplete rung between them (see Table 6.1 and Fig. **??**). We used the same reaction coordinates as those used previously for the GB1 β-hairpin [202] (see Methods). Upon termination of the SLUS procedure, 15 ns/window of replica exchange umbrella sampling (REUS) was performed on all generated windows. The resulting 2D folding PMFs are shown in Fig. 6.1.

The native state for each band is unstable in isolation. In some cases – namely, bands 1C and 3N – the native state was not captured by the SLUS procedure. For these bands, the native state was quickly lost within the first nanosecond of the initial 310-K equilibrium simulations, and the energetic barrier to refold them was too large for the SLUS procedure to overcome with only a 5 kcal/mol cutoff. For the remaining bands, only 3C and 2N possess a free energy minimum near the native state. Band 1C is the only other band to possess a minimum outside of the unfolded regime, with two minima at $\sim$3.5 and $\sim$8.5 hydrogen bonds, with all minima only separated by barriers of $<$2 kcal/mol. Band 2N also exhibits some intermediate folded states between 5-8 hydrogen bonds, with minima separated by barriers of $<$3 kcal/mol.

It is clear from our PMFs that without a scaffold of neighboring structures, the bands of the β-helix cannot form on their own, or do so very infrequently. The two terminal bands, 1C and 2N (band 1N most likely serves as a "cap" to prevent aggregation of the β-helix [203]), exhibit the highest stability of all the bands we tested. Band 1C exhibits

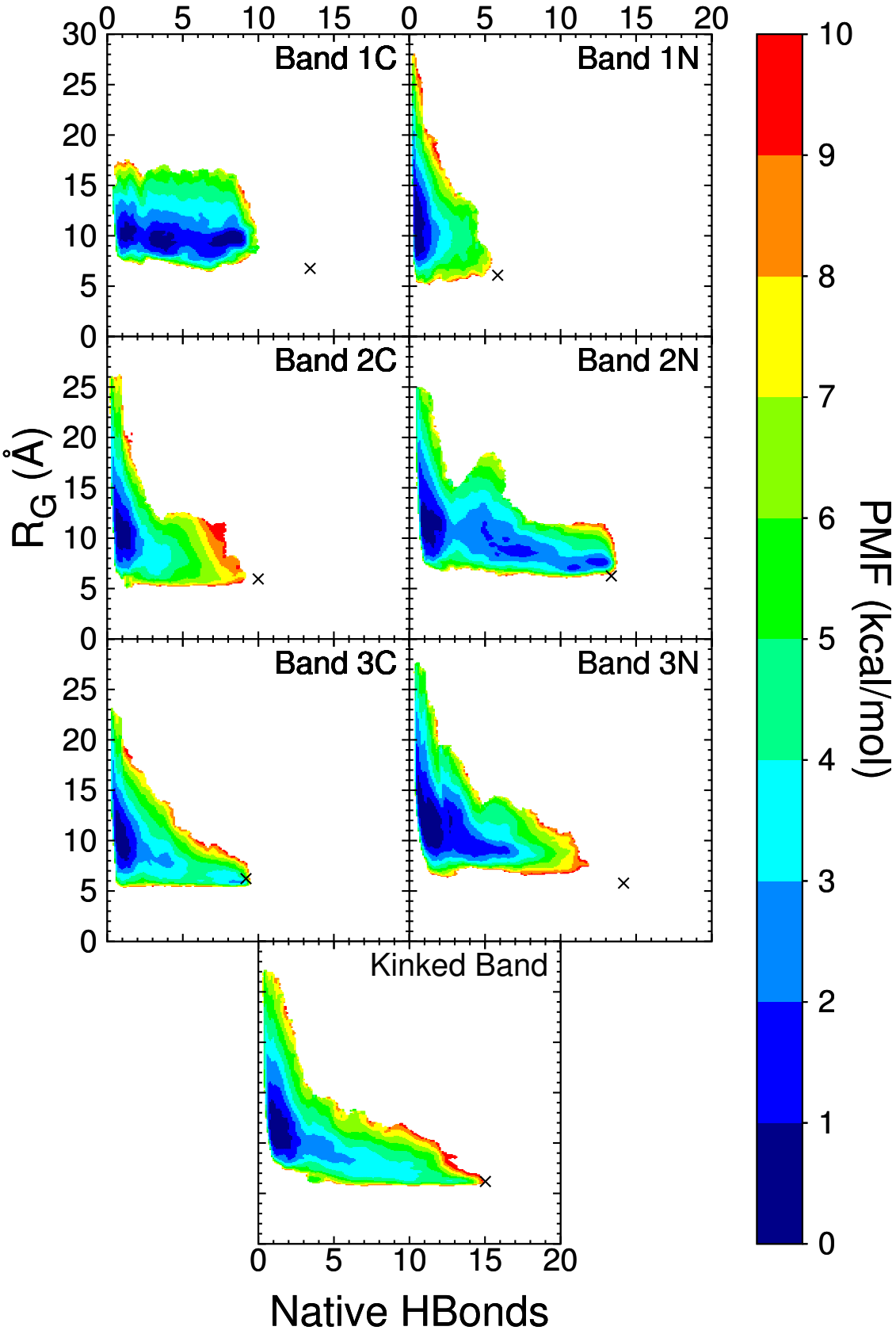


Figure 6.1: *Folding PMFs for isolated bands of the pertactin β-helical passenger domain.* Crystal structure coordinates are plotted with an "x".

three states that exist in roughly equal proportions – one unfolded state and two partially folded states. If we combine the two partially folded states to one single state, then we get $\Delta G_{\text{fold}} = -0.80 \pm 0.48$ kcal/mol, indicating that more often than not – $\sim$79% of the time – some native structure is present for this band. If we restrict it to only the most structured minimum ($\sim$60% folded), these values reduce to $+0.15 \pm 0.39$ kcal/mol and $\sim$44%, respectively. Band 2N also exhibits three minima – one unfolded, one partially ($\sim$40-60%) folded, and the fully folded native state. Combining the two folded states, we get $\Delta G_{\text{fold}} = +0.68 \pm 0.62$ kcal/mol, with native structure present $\sim$25% of the time. Considering only the native state minimum, these decrease to $+1.67 \pm 1.07$ kcal/mol and $\sim$6%, respectively. The band with the next highest percent of observed native structure is band 3C at only $\sim$0.7%. All other bands are $<$0.1%.

Given the much higher levels of native structure at the termini relative to the rest of the passenger domain, bands 1C and 2N may act as scaffolds for the propagation of β-helical structure inward towards the kinked region, giving rise to the concerted folding process observed for the pertactin passenger domain *in vitro* [7]. Both termini exhibit at least 50% native structure at about the same rate, which could explain why the folding rates are the same for the N- and C-terminus. Meanwhile, the scarcity of full native structures in any band may account for the extremely slow folding rate for the overall structure [6].

6.3.2 Vectorial folding stabilizes the β-helical bands

Next, we calculated the C$\rightarrow$N vectorial folding free energy landscapes for the same bands tested in the previous section. These landscapes were calculated by restraining all C-terminal β-sheets to their native conformations while only allowing the most N-terminal rung to move freely (see Methods for more details). PMFs for bands 2C and 1N are shown in Fig. 6.2.

While these bands 2C and 1N were completely unstable when isolated (see Fig. 6.1), they become very stable when forming on top of previously formed C-terminal bands. The

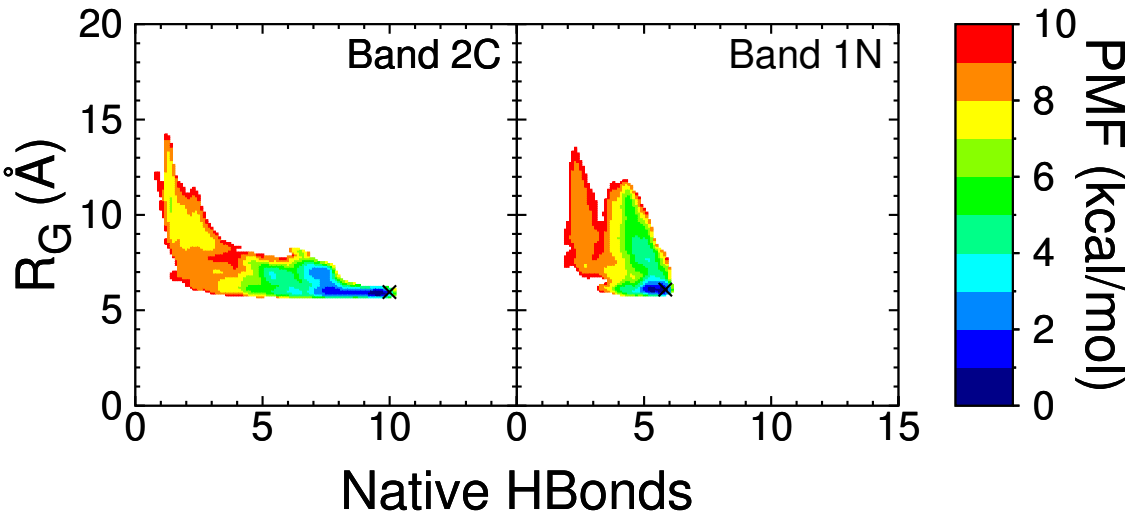


Figure 6.2: *C→N vectorial PMFs for bands of the pertactin β-helical passenger domain.* Crystal structure coordinates are plotted with an "x".

sharp drop in entropy for the unfolded conformations, particularly for the C-terminal rung of each band which is heavily restrained, contributes to large increase in energy for unfolded state. If we assume the folded states are roughly the same, the unfolded states are ∼20 kcal/mol higher in the vectorial system than in the isolated system for these two bands. In the vectorial folding hypothesis, during OM secretion, only small sections of the peptide immediately N-terminal to the already secreted and folded bands of the helix are presumed to also be in extracellular space, whereas on the periplasmic side the remaining peptide wriggles freely. The entropy of the periplasmic portion is presumably even higher than for our isolated systems, so the increase in the free energy of the unfolded state could be significantly larger than observed here, further stabilizing the folded states on the extracellular side. The vectorial pathway no only prevents backstracking of the peptide through the β-domain, but it eliminates the barrier for folding and significantly stabilizes the the folded state in the process.

6.4 Discussion

We used self-learning adaptive umbrella sampling (SLUS) and replica exchange umbrella sampling (REUS) to examine and compare the *in vitro* and *in vivo* folding pathways of the β-helical passenger domain of the pertactin autotransporter. From our 2D folding PMFs we observed that the vectorial pathway significantly stabilizes the β-sheet bands of the β-helix and greatly increases the folding rate over the concerted, *in vitro* pathway. Bands 2C and 1N, which showed no propensity for forming native β-sheets when isolated from the rest of the passenger domain, were completely stable when the previous C-terminal band was present and in its native conformation. The vectorial pathway stabilized these bands by $\sim$20 kcal/mol, inducing cooperative folding of the β-helix. In addition, bands 1C and 2N are the only bands that exhibit any significant amount of native structure in isolation, with at least half of the native hydrogen bonds present 44% and 25% of the time, respectively. No other band of the helix is above 1%, with most less than 0.1%. The two terminal bands may act as scaffolds to induce folding of the internal bands *in vitro*, leading the observed concerted pathway [6, 7]. The overall lack of native structure observed in our isolated systems could also explain why folding is so slow along this pathway.

CHAPTER 7
CONFORMATIONAL DYNAMICS OF ACRA GOVERN MULTIDRUG EFFLUX PUMP ASSEMBLY

7.1 Introduction

The development of antibiotic resistance in pathogenic bacteria is quickly becoming one of the most dangerous global pandemics that has been observed in modern times, responsible for approximately two million illnesses and over 23,000 deaths annually in the Unites States alone according to a recent report from the CDC [13]. Some strains of *Pseudomonas aeruginosa* are now resistant to all classes of antibiotics, and *Escherichia coli* is not far behind, with some strains resistant to at least six classes of antibiotics [14]. One of the primary sources of this resistance is in the recognition and expulsion of antimicrobial drugs from the cell via large, cellular envolope-spanning multidrug efflux pumps [14].

The efflux pump in *E. coli* is the tripartite AcrAB-TolC complex [204]. Structural studies show it is composed of three main components, AcrB, AcrA, and TolC, in a 3:6:3 ratio [15, 205, 206]. Substrate recognition and energy transduction are performed by the AcrB trimer at the inner membrane [207, 208], while the TolC trimer forms a pore in the outer membrane. The membrane fusion protein (MFP), AcrA, which assembles as a trimer of dimers, completes pump assembly by binding together the inner and outer membrane components. So-called efflux pump inhibitors (EPIs) are emerging alternative therapeutics that have the potential to revive activities of existing antibiotics and to control the spread of antibiotic resistance [209]. While earlier studies have focused on inhibiting the transport activity of AcrB [209, 210], new studies have begun focusing on the MFP, AcrA, in an effort to diverse to pool of potential drugs [16, 17].

AcrA has four main structural components: the α-hairpin, lipoyl, β-barrel, and mem-

brane proximal (MP) domains [207]. The hairpin domain interacts with the outer membrane channel, TolC, while the other domains interact with AcrB [211, 212, 213, 214]. Previous molecular dynamics simulations of a free AcrA monomer reveal that while the two central domains, the lipoyl and β-barrel domains, are rigid, the two terminal domains, the hairpin and MP domains, are quite flexible [215]. The question then arises of how this flexible monomer binds to AcrB and TolC as a dimer. In this work we wish to better understand the flexibility of AcrA and how it impacts binding to AcrB. We employ self-learning adaptive umbrella sampling (SLUS) [8] with molecular dynamics simulations to characterize two main conformations of AcrA, as well as how these conformations change upon binding to AcrB.

7.2 Methods

7.2.1 Self learning adaptive umbrella sampling (SLUS) of free AcrA monomer

We used the model of free AcrA developed by Wang and coworkers as a starting conformation [215]. This model is composed of the crystal structure for the α-hairpin, lipoyl, and β-barrel domains (PDB code: 2F1M), and a homology model of the membrane proximal (MP) domain based on the structure for MexA (PDB code: 2V4D) (see Fig. 7.1). The protein was solvated in water box of size 201 $\times$ 81 $\times$ 124 Å^3 with 62,402 TIP3P water molecules [92], and neutralized with 0.15 M NaCl, for a total system size of 192,773 atoms. A harmonic restraint was added to align the lipoyl–β-barrel vector along the x-axis. Simulations were performed using NAMD 2.10-12 [**Phillips2005**] with the CHARMM36m protein force field [184]. The temperature was fixed at 300 K using Langevin dynamics; the pressure was kept constant at 1 atm using the Langevin piston method [96]. The equations of motion were integrated using the RESPA multiple time-step algorithm with a time step of 2 fs used for all bonded interactions, 2 fs for short-range non-bonded interactions, and 4 fs for long-range electrostatic interactions. Long-range electrostatic interactions were calculated using the particle-mesh Ewald method with a real-space cutoff of 12 Å [97].

Short-range non-bonded Lennard-Jones interactions were cutoff at 12 Å with a potential switching function beginning at 10 Å bringing the potential energy smoothly to zero at the cut off distance. Bonds involving hydrogen atoms were constrained to their equilibrium length, employing the SETTLE algorithm [146] for water molecules and the SHAKE algorithm for all others [147].

We ran the self-learning adaptive biasing procedure [8] starting from a set of 27 initial windows centered around (θ,ϕ,ψ) = (120°,115°,155°) generated from a 1-ns equilibrium simulation. For our collective variables, we use the three angles previously defined by Wang et al [215], namely the angles the two terminal domains make with the two central domains (θ for the α-hairpin domain and ψ for the MP domain) and the dihedral angle between these domains about the central axis (ϕ; see Fig. 7.1). Windows were spaced by 5° with a 0.2 kcal/mol force constant for all three angles. Each window was simulated for 2 ns, and PMFs were generated using the weighted histogram analysis method [45]. A modified version of the WHAM code from Alan Grossfield [155] was used to generate 3D PMFs. We also implemented the method of direct inversion in the iterative subspace (DIIS) to accelerate the convergence [59].

A 3 kcal/mol free energy cutoff was used to generate new windows in each iteration of the SLUS procedure. Initially, the procedure halted near ϕ = 150° after 18 iterations resulting in a total of 743 windows, upon which we then ran 3 ns/window of replica exchange umbrella sampling (REUS) [51] on all windows. After the REUS simulations, the SLUS procedure was then allowed to proceed up to ϕ = 240°, generating an additional 830 windows after another 17 iterations, for a total of 1573 windows and 35 iterations. We then ran a final REUS simulation on all 1573 windows for 15-ns/window, and generated a final 3D PMF. Total simulation time was 26.7 μs.

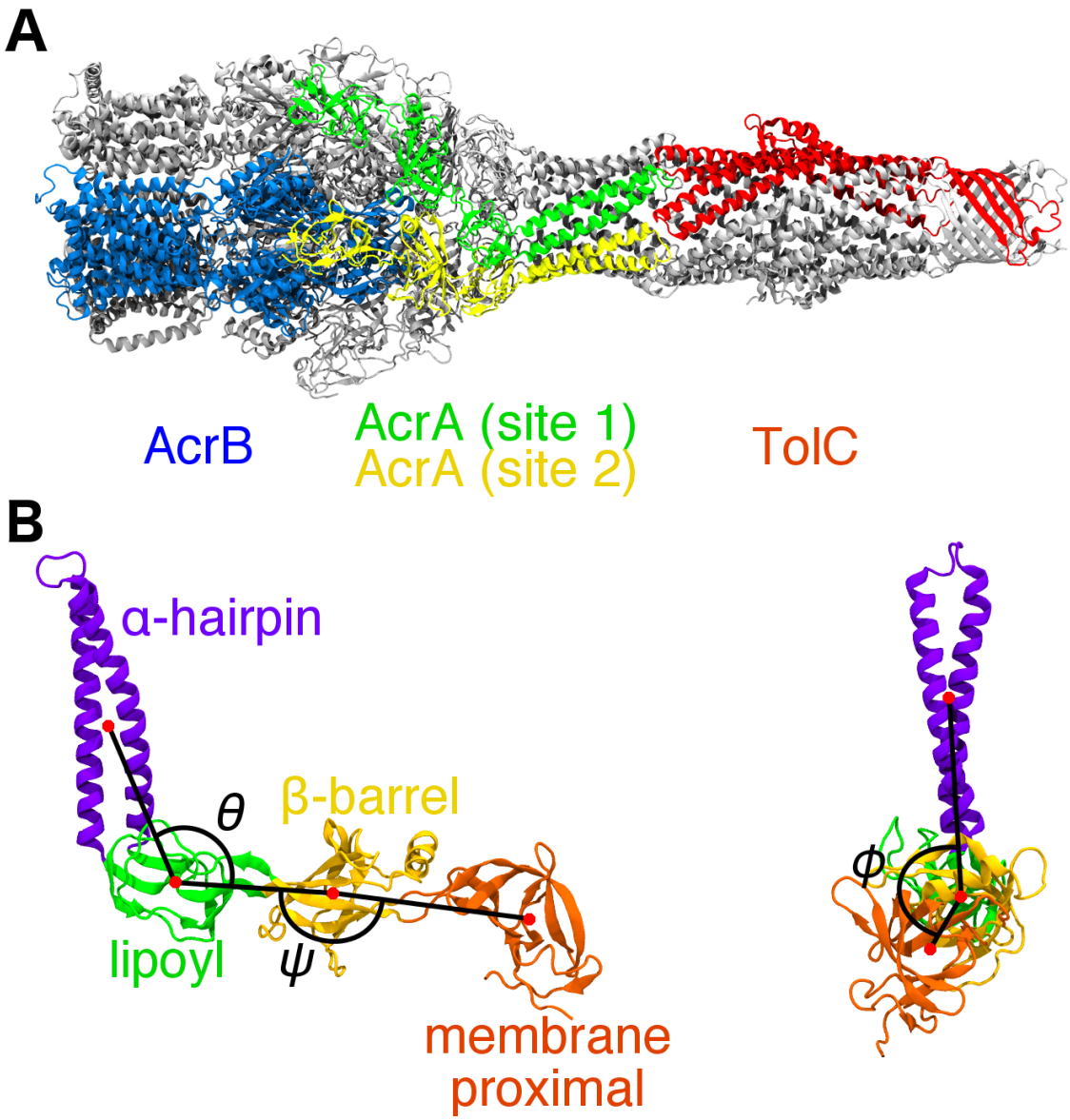


Figure 7.1: *The AcrAB-TolC efflux pump complex.* (A) Cryo-EM structure of the full AcrAB-TolC efflux pump complex from Ref. [206]. Entire structure shown in cartoon representation. One copy of the inner membrane transporter, AcrB, is colored blue, and one copy of outer membrane channel, TolC, is colored red. Two copies of the membrane fusion protein, AcrA, in binding sites 1 and 2 are shown in green and yellow, respectively. All other copies of AcrA, AcrB, and TolC are colored gray. (B) Structural domains of AcrA shown in a cartoon representation, with the four domains colored and labeled. The three angles, θ, ϕ, and ψ, previously defined in Ref. [215] are also labeled. (Left) Side view. θ is the angle between the α-hairpin, lipoyl, and β-barrel domains. ψ is the angle between the lipoyl, β-barrel, and membrane proximal (MP) domains. (Right) Front view. ϕ is the dihedral angle between the α-hairpin and MP domains along the axis defined by the lipoyl and β-barrel domains.

7.2.2 Equilibrium simulations of free AcrA monomer

For the starting structure of AcrA, we used one of the bound conformations from Ref. [216]. The protein was then solvated with 156,554 TIP3P water molecules [92] in a cubic box of length 168 Å, and then ionized with 0.15 M NaCl, for a total system size of 475,761 atoms. The CHARMM36m [184] force field parameters were used to describe the protein. Hydrogen mass repartitioning [217] was used for all protein atoms in order to utilize of 4-fs timestep. A 12-Å Lennard-Jones cutoff with switching beginning at 10 Å was used for van der Waals interactions. Simulations were performed using Amber 16 [218] on GPUs. The temperature was held constant at 300 K using a Langevin thermostat. Periodic boundary conditions were used, and the pressure was held constant at 1 atm using the Langevin piston method [96]. The particle mesh ewald (PME) method was used to describe long-range electrostatics [97]. Bonds involving hydrogen atoms were constrained to their equilibrium length, employing the SETTLE algorithm [146] for water molecules and the SHAKE algorithm for all others [147]. We first ran a pre-equilibration run of 1 ns of WT AcrA, after which AcrA was mutated. Finally, we ran 250-ns production runs of the WT and mutated systems. We also ran three additional simulations of the unbound WT AcrA model that was used to initiate the SLUS procedure. The same procedures described for the bound conformation were applied to the unbound conformation.

7.2.3 SLUS of AcrB-bound AcrA

The AcrB trimer and one copy of bound AcrA were extracted from the cryo-EM structure of Jeong et al. [216]. We built two systems, one with a single copy of AcrA in binding site 1, and another with a single copy in binding site 2. All other binding sites were left unoccupied. The AcrB trimer was embedded within a symmetric bilayer composed of 924 total lipid molecules, with the following lipid composition: 409 PMPE, 110 POPE, 110 QMPE, 84 PMPG, 82 PYPG, 74 YOPE, and 54 PVCL2 lipids. The N-terminus of AcrA was lipidated with a POPE lipid tail, which was also embedded in the bilayer. Each

system was then solvated in a water box of size of 190 $\times$ 190 $\times$ 256 Å^3 with $\sim$243,000 TIP3P water molecules [92], and neutralized with 0.15 M NaCl, for a total system size of $\sim$903,00 atoms. Simulations were performed using NAMD 2.10-12 [**Phillips2005**] with the CHARMM36m protein force field [184] and the CHARMM36 lipid force field [219].

For each system, we ran a 1-ns equilibrium simulation to seed our SLUS procedure. From this 1-ns simulation, we also extracted a set of contacts between AcrA and the AcrB trimer. We tracked when a residue on AcrA came within 5 Å of a residue on the copies of AcrB. For all such residue pairs, we measure the mean and standard deviation of their separation distance. For our system, AcrA is in contact with two of the copies of AcrB, so we picked the five residue pairs between AcrA and both of these copies of AcrB with the lowest standard deviation and restrained them to their mean distance with a 2.5 kcal/mol·Å^2 force constant. With these restraints, the ψ angle between the MP and β-barrel domains is severely restricted. Therefore, for our US simulations, we only biased the θ and ϕ angles.

The SLUS procedure proceeded similarly to the free AcrA system. For site 1, we started with 6 initial windows, and after 25 iterations, the SLUS procedure produced a total of 269 windows. For site 2, we started with 28 initial windows, and after 47 iterations, the SLUS procedure produced a total of 411 windows. These were both followed by REUS simulations of 15 ns/window, for a total simulation time of 4.6 μs and 7.0 μs for site 1 and site 2, respectively. 2D PMFs were then generated using WHAM [155].

7.2.4 Experimental protocols

For chromosomal insertion of WT and AcrA mutants, lambda red homologous recombinant technique was used, selection of insertion was confirmed by colony PCR and AcrAB substrates counter screening. A mini-Tn7T-based protocol was used to insert the pore (fhuAΔC/Δ4L) into the *E. coli* chromosome using the pUC18T-R6K-mini-Tn7T suicide delivery vector along with the pTNS3 helper plasmid. The insertions were confirmed by PCR and vancomycin disc susceptibility assay. The vancomycin disc assay was described

earlier in detail. [220]

7.3 Results

7.3.1 3D PMFs of free AcrA monomer calculated from self-learning umbrella sampling

Using self-learning umbrella sampling (SLUS, see Methods), we explored the conformational space of free AcrA monomer in solution using the three angles, θ, ϕ, and ψ, first defined by Wang et al. [215] and shown in Fig. 7.1B. Starting from an initial set of 27 windows generated from a 1-ns equilibrium simulation, we ran 35 iterations of the self-learning algorithm at 2 ns/window with a 3 kcal/mol free energy cutoff, to generate a total of 1573 windows that spanned ranges of 45°, 210°, and 60° for θ, ϕ, and ψ, respectively. The self-learning scheme extends the range of sampled conformations over previously reported short (20 ns) simulations of free AcrA monomers, which only covered ranges of 25°, 180°, and 60°, respectively. [215] A 3D PMF was then calculated using replica-exchange umbrella sampling (REUS, see Methods). In total, 26.7 μs of simulation time was used to calculate the full conformational free energy landscape of AcrA. The full 3D PMF, as well as corresponding 2D and 1D PMFs calculated by integration of the 3D PMF along all three angles (see Methods) is shown in Figure 7.2.

As predicted from previous simulation studies, free AcrA is quite flexible in the absence of AcrB and TolC, exploring a wide range of orientations of its four structural components. Our results reveal a previously unreported energetic barrier separating two main conformational basins. As can be seen more clearly from the 1D PMFs, this barrier lies mostly along the dihedral angle, ϕ, around 150° and is relatively small ($\sim$1-2 kcal/mol). The high-ϕ, *trans*-like conformational basin is $\sim$1 kcal/mol deeper than the low-ϕ, *cis*-like conformation, but is also narrower along ϕ by $\sim$45°. Conversely, the *trans* conformation exhibits a larger range in ψ, extending $\sim$20° lower than the *cis* conformation. Ultimately, the two conformations are roughly equivalent in free energy, with $\Delta G_{cis \rightarrow trans} = G_{trans} - G_{cis} = -0.21 \pm 0.18$ kcal/mol, or about a 40/60 *cis*/*trans* equilib-

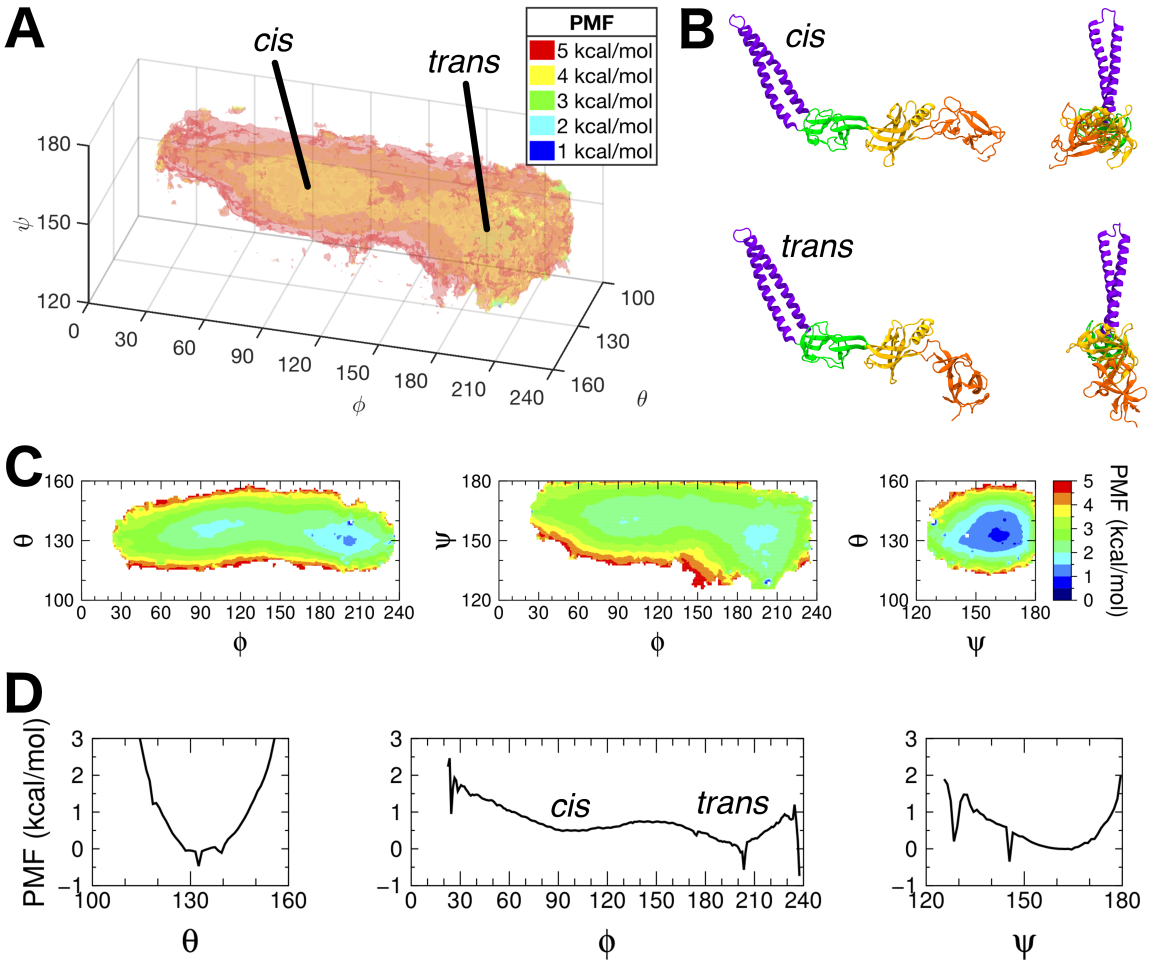


Figure 7.2: *Multidimensional PMFs of free AcrA monomer in water.* (A) Free energy landscape (PMF) of free AcrA in water along the θ, ϕ, and ψ angles. The 3D PMF was calculated using replica-exchange umbrella sampling (REUS) with a total 1573 windows simulated for 15 ns/window. Isosurfaces of 0.5 kcal/mol are plotted. Two free energy basins – *cis* and *trans* conformations – are separated by a small energy barrier. (B) Cartoon representations of the two conformations of AcrA, with side and front facing views. (C) 2D PMFs calculated by integrating out each angle of the 3D PMF. (D) 1D PMFs calculated by integrating out a second angle from the 2D PMFs. See Methods for more detail on how each PMF was calculated.

rium ratio. Cryo-EM structures of the full AcrAB-TolC complex [206, 216] show AcrA to be in the *trans* conformational basin, near the top ($\psi > 160°$; see Table 7.1 and Fig. 7.1A).

Table 7.1: *Conformations of AcrA within the full AcrAB-TolC efflux pump complex.* Three AcrAB-TolC structures were examined: one with an open conformation of the outer membrane channel, TolC (Ref. [216]), and two with a closed conformation of TolC (Ref. [206]). Each structure has three copies of AcrA in both site 1 and site 2, for a total of 9 copies for each site. Averages and standard deviations for each site are tabulated below.

	θ (°)	ϕ (°)	ψ (°)
Site 1	124.4 ± 4.3	181.7 ± 8.0	160.7 ± 1.1
Site 2	119.6 ± 1.5	179.5 ± 8.9	168.7 ± 2.4

While the *cis* conformational basin is mostly smooth, several small minima occur throughout the *trans* conformational basin, particularly on the boundaries of the PMF in the low-ψ ($<150°$) region, such as those near (θ,ϕ,ψ) = (140°,205°,130°) and (132°,245°,145°), which are actually the global minima of the PMF. These can be seen more easily in the 1D PMFs (see Fig. 7.2D). The depth of these minima also changes significantly with simulation time, changing by 1-3 kcal/mol over the course of the REUS simulations (see Fig. SXX). However, due to their small size compared to the overall basin, their contribution to the overall free energy is minimal, with block averaging producing only a small error of 0.18 kcal/mol (see Methods).

7.3.2 Transition state between *cis* and *trans* conformations involves interactions between MP domain and α-helix of β-barrel domain

The free energy landscape of free AcrA is composed of two large conformational basins, *cis*- and *trans*-like conformations, separated by a small energy barrier. A transition state lies in the narrow pathway connecting the two basins between ϕ = 120°-150°. We examined the interactions which distinguish the two conformations and the changes that occur at the transition state. We ran four additional 250-ns equilibrium simulations of free AcrA, three starting from the *cis* conformation and one starting from an AcrAB-TolC bound conformation, to augment our REUS simulations.

While in the *cis* conformation, the α-helix of the β-barrel domain forms contacts with the MP domain in two main regions: (1) the β-hairpin consisting of residues 339 to 350 and (2) the connector strands between the MP and β-barrel domains consisting of residues 48 to 53 and 299 to 307 (see Fig. 7.3A and B and Table 7.2). These two regions combine in roughly equal parts to form the interfacial site which governs the flexibility of the *cis* conformation. In the transition region observed in our 3D PMF ($120^\circ < \phi < 150^\circ$), we see a clear decline in the number of contacts in both the hairpin and connector regions as the protein changes from the *cis* to the *trans* conformation (see Fig. 7.3C). We calculated the occupancy of these contacts by Boltzmann reweighting each trajectory frame of our REUS simulations according to AcrA's free energy; i.e., each frame is weighted by $\exp[-W(\theta, \phi, \psi)/kT]$, where $W(\theta, \phi, \psi)$ is the 3D PMF, T is the temperature, and k is the Boltzmann constant. From this, we observe a similar decline in contacts during the *cis*-to-*trans* transition (see Fig. 7.3D-F). In addition, from our equilibrium simulations, we see that the conformational space not sampled in our REUS simulations ($\phi > 240^\circ$) roughly mirrors the previously described *cis* conformation, reforming interfacial contacts between the MP β-hairpin and the β-barrel α-helix. Based upon our 3D PMFs, it's possible that $\phi > 240^\circ$ could be a second *cis* conformational state, energetically distinct from the first *cis* conformation ($\phi < 150^\circ$). For now, we will simply refer to both as the *cis* conformation.

Residues Arg225, Gln228, and Asn232 on the β-barrel α-helix compose the majority of the contacts with the MP β-hairpin, making contacts primarily with residues Gln341 and Ile343, with Arg225–Ile343 and Gln228–Gln341 being two of the most occupied contacts in both our equilibrium and REUS simulations. Residues Arg225 and Glu229 form the majority of the contacts with the MP–β-barrel connector strands, forming contacts with Leu50, Gln51, Ile52, Glu299, and Asn302, with Glu229–Gln51, Glu229–Asn302, and Arg225–Gln51 three of the most occupied contacts with the connector. This combination of residues, along with additional contacts involving residues Asn221, Ser340, Ala342, Leu348, and Thr350, creates an amphiphilic cluster, with charged, polar, and hydrophobic

Table 7.2: *Contacts at the interfacial site between the MP and β-barrel domains in the* cis *conformation.* The *cis* conformation is defined as $\phi < 150°$ and $\phi > 240°$. Two residues are considered in contact if their side chains are within 5 Å of each other. Ten highest occupied contacts in both the equilibrium and REUS simulations, the latter reweighted by to the free energy, are tabulated below.

β-hairpin				Connector			
Residues		% Occupied		Residues		% Occupied	
β-Barrel	MP	Equilibrium	REUS	β-Barrel	MP	Equilibrium	REUS
R225	I343	35.67	79.15	R225	Q51	29.17	50.85
Q228	Q341	37.68	66.97	E229	N302	29.22	49.96
R225	G344	7.96	49.97	E229	Q51	42.57	41.55
Q228	I343	12.53	41.14	R225	E299	0.37	33.03
N232	Q341	40.15	38.59	R225	I52	24.33	30.99
L224	I343	15.59	37.85	E229	E299	8.37	29.30
R225	Q341	23.14	37.82	R225	L50	27.66	27.91
N221	I343	17.88	36.79	E229	G300	6.17	23.14
Q228	S340	28.03	36.57	L226	E299	10.70	22.08
R225	A342	7.27	35.15	R225	T53	4.44	20.73
N232	S340	34.02	34.47	E229	I52	38.23	6.65
N232	I343	32.42	10.75	E229	L50	25.65	19.96
N232	L348	30.50	25.12	R226	I52	23.50	2.12
N232	T350	24.01	14.25	N232	L50	16.83	8.99
L224	Q341	23.46	33.95	L226	Q51	15.45	7.14

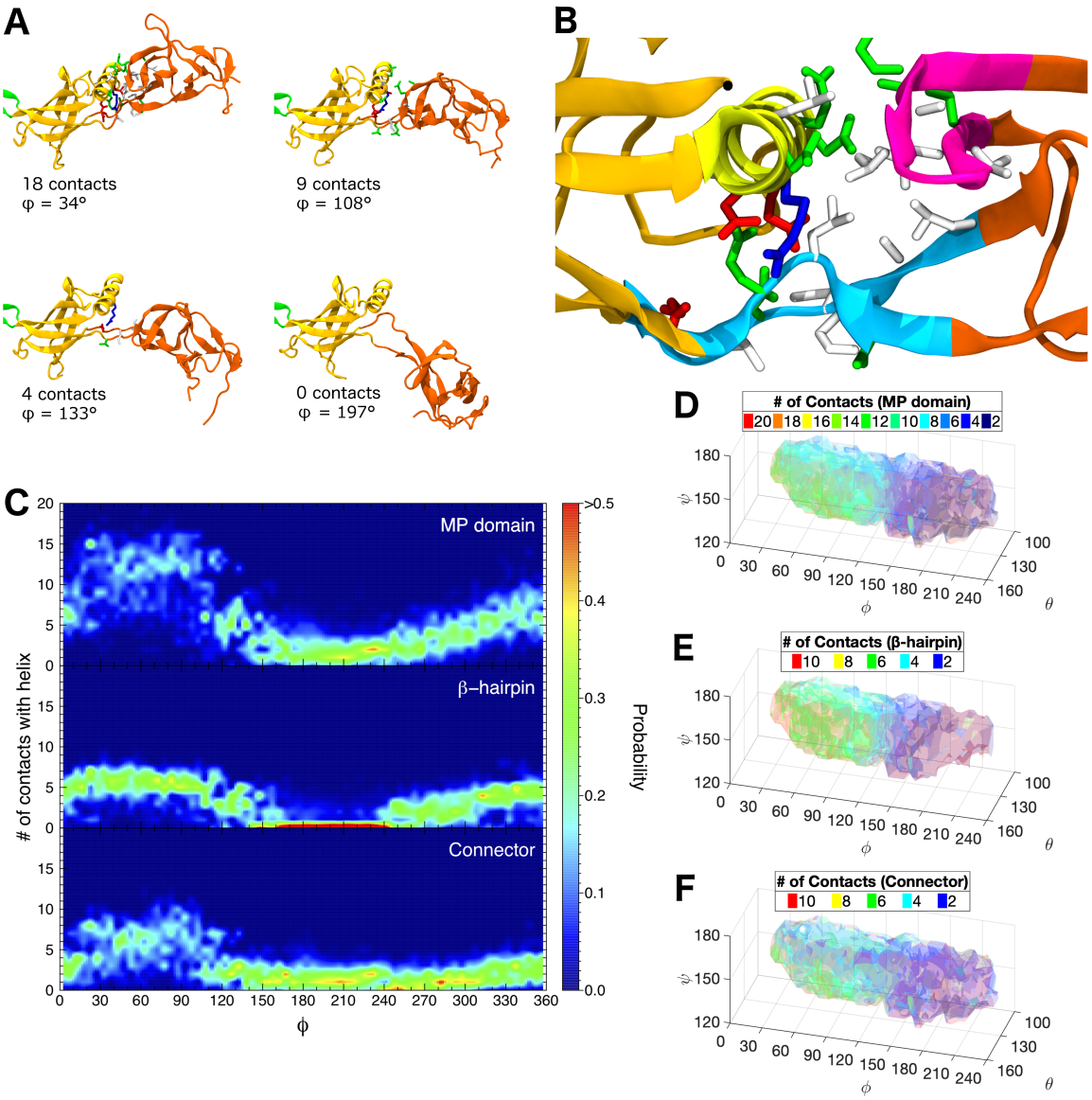


Figure 7.3: *Interactions between β-barrel and MP domains distinguish the two main basins of the conformational landscape of free AcrA monomers.* (A) Snapshots from a 250-ns equilibrium simulation of free AcrA starting from the *cis* conformation. Protein backbone is shown in cartoon representation, with MP domain colored red, β-barrel domain colored orange, and lipoyl domain colored green. Contact residues on the MP domain and the α-helix of the β-barrel domain are shown in licorice representation with hydrogen atoms omitted, colored by residue type: (blue) positively charged, (red) negatively charged, (green) polar, and (white) hydrophobic. Two residues are in contact if their side chains are within 5 Å. The number of contacts is defined as the number of MP domain residues. Number of contacts and ϕ values are denoted below each conformation. (B) Close up of the interfacial site between the β-barrel domain and the MP domain. Helical residues 220 to 232 are colored yellow, MP β-hairpin residues 339 to 350 are colored magenta, and connector residues 48 to 53 and 299 to 307 are colored light blue. (C) Distribution of the number of residues of the MP domain in contact with β-barrel helix over the dihedral angle, ϕ, for (top) the entire MP domain, (center) the β-hairpin, and (bottom) the connector. (D)-(F) Number of MP domain residues in contact with the β-barrel helix calculated from our REUS simulations for (D) the entire MP domain, (E) the β-hairpin, and (F) the connector.

residues forming a wide range of interactions. The small energy barrier going from the *cis* to the *trans* conformation, despite the large number of contacts that must be broken during the transition, could be due to the lack of a well-defined hydrophobic cluster, salt bridge, or other strong interaction holding the interfacial site together. Mutations at the interfacial site that enhance either the hydrophobicity or hydrophilicity could significantly alter the *cis*/*trans* balance by strengthening or weakening, respectively, the stability of the MP–β-barrel interface. For bound conformations of AcrA within the full efflux pump complex, only residues Arg225, Glu229, Ser340, and Gln341 at the interfacial site form contacts, with Arg225–Gln341 being the most prevalent among all the structures, and these contacts are only observed in one of the binding sites. This is consistent with AcrA primarily being in the *trans* conformation in the available structures (see Table 7.1). Given this preference for the *trans* conformation and the lack of interfacial contacts within the assembled complex, modifying the interfacial site could significantly affect the efficiency of pump assembly and/or function.

7.3.3 Mutational analysis of AcrA interfaces

MD simulations have identified contacts between the MP and β-barrel domains that distinguish two conformations of free AcrA. To study the importance of these contacts to pump assembly and function, we used site-directed mutagenesis to introduce substitutions that disrupt them. The AcrA residues Leu50, Ile52, Thr53, Arg225, Glu299, Ile343, Lys346, and Trp347 in the MP domain and the interface between the MP and β-barrel domains were substituted with alanine one at a time. At the hairpin and lipoyl domain interface, the previously described residues Glu67, Arg69, Thr174, Lys175, and Thr177 substituted with alanine, arginine (E67), or tryptophan (E67) [16] were included for comparison.

All AcrA variants were produced at the levels comparable to the wild type protein (see Fig. 7.4) and complemented the antibiotic susceptible phenotype of *E. coli* ΔacrAB cells (Table 7.3) except for L50A (interfacial), which could be due to improper folding. In

agreement with previous studies [16], the activity of the pump with AcrA E67R (hinge) was reduced by 2-4 fold. Similarly, cells carrying AcrA I52A (interfacial) variants were 2-4 fold more susceptible to antibiotics than cells producing the wild type AcrA.

Table 7.3: *Antibiotic susceptibility of E. coli WT-Pore cells carrying the plasmid borne AcrAB with indicated AcrA variants.*

	No Pore				Pore			
	NOV	ERY	TET	SDS	NOV	ERY	TET	SDS
–	1	0.5	0.25	19.5	0.5	0.25	0.06	9.7
pUC18	1	1	0.25	39	0.12	0.12-0.2	0.06	9.7
WT	256	64	0.5	>10000	32	8	0.5	312
I52A	64	16	0.25	>10000	4-8	1-2	0.12	78
T53A	128	32	0.25	>10000	16	2-4	0.25	78-156
R225A	128	32	0.25	>10000	16-32	4-8	0.25	78
E299A	128	32	0.25	>10000	8-16	8	0.25	78
I343A	128	32	0.25	>10000	8-16	2	0.12	39-78
I343W	256	64	0.5	>10000	8	2-4	0.12	39-78
K346A	256	32	0.25	>10000	32	8	0.25	78-156
E67A	128	32	0.25	>10000	8	2-4	0.12	39-78
E67R	128	16	0.25-0.5	>10000	4-8	1-2	0.12	39
R69A	256	32	0.25-0.5	>10000	8-16	8	0.25	78
T174A	256	32	0.25	>10000	8	2	0.12	39-78
K175A	256	32	0.5	>10000	4-8	4	0.25	78
T177A	256	64	0.5	>10000	32	8-16	0.5	156-312

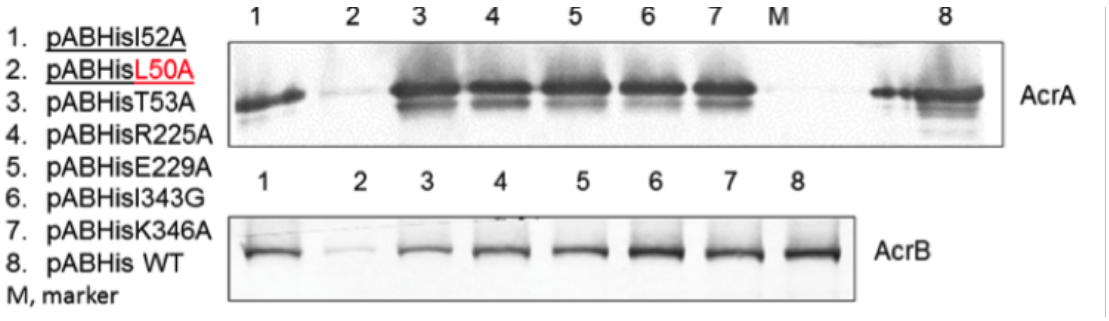


Figure 7.4: *Expression of AcrA variants and the co-expressed AcrB.*

Since AcrAB-TolC activities are strongly affected by the low permeability barrier of the outer membrane, we also analyzed the functionality of the plasmid-borne efflux pumps assembled with AcrA variants in hyperporinated ΔacrAB(Pore) cells. Among tested antibiotics the potencies of erythromycin and SDS dropped significantly upon hyperporination,

even in cells overproducing the wild type AcrAB-TolC (Table 1). In contrast, the potencies of novobiocin and tetracycline were only modestly, if at all, affected by hyperporination (see Table 1). Thus, the potencies of novobiocin and tetracycline, but not of erythromycin or SDS, are largely limited by the activity of AcrAB-TolC. MIC measurements showed that AcrA mutants containing substitutions of Glu67, Thr174 and Lys175 at the hinge site and Ile52 and Ile343 at the interface site were only partially active in ΔacrAB(Pore). Thus, substitutions in these positions at the hinge and interfacial sites of AcrA negatively affect the functionality of the assembled pump.

To confirm that the observed phenotypes are not sensitive to the mode of expression, the acrAB operons producing mutated AcrA variants were integrated into the native location on the *E. coli* chromosome. The MICs of antibiotics were 2-4 fold higher in cells carrying the plasmid-borne acrAB than in cells with chromosomal expression of acrAB and its mutated variants, but in agreement with the susceptibility profiles of cells carrying acrAB variants on plasmids (see Table S1). Hence, the changes in activities of AcrA variants are intrinsic to the protein and not to its mode of expression.

7.3.4 Mutations at the interfacial site alter the flexibility of AcrA

Several mutations at the interface between the MP and β-barrel domains of AcrA were shown to reduce the efficiency of the AcrAB-TolC pump, particularly I343A and I52A. We ran additional 250-ns equilibrium simulations of free AcrA in water starting from a bound conformation with the experimentally tested mutations and compared the flexibility to that of the wild type. We also ran two mutants near the hinge site between the α-hairpin and lipoyl domains, E67A and E67W, for comparison. The conformational distributions of the WT and mutant simulations are shown in Fig. 7.5 and Supplemental Fig. SXX. While we expect there to be two conformational basins based on our free energy landscapes, our WT simulation starting from the bound state remained in the *trans* conformation for the entire 250 ns, so the simulation time used here may be insufficient for discerning *cis*/*trans* distri-

butions. For example, the I343W, E67W, and T53A systems also remained in the WT *trans* conformation, while the E67A and T53A systems remained in the WT *cis* conformation as expected from our free energy calculations (see Fig. 7.2). However, we can discern extreme changes in the conformational distribution or to the energy barrier between the two conformations. One common characteristic of all the systems, for example, was that the α-hairpin domain was far from the bound conformations in the full pump complex. The presence of the two transmembrane domains, AcrB and TolC, may influence the conformation of the α-hairpin domain of AcrA as part of the assembly process. We also tracked the contributions to the dihedral angle, ϕ, from the α-hairpin and MP domains, and found that motions along ϕ in all cases are dominated by the MP domain (see Fig. 7.5, bottom graph), signifying the importance of the MP domain in the conformational distribution of AcrA.

Arg225 and Ile343 are two critical residues at the interfacial site as they contribute significantly to its amphiphilic nature. The R225–I343 contact is also one of the most highly occupied between the β-barrel α-helix and the MP β-hairpin in the *cis* conformation (see Table 7.2). Consequently, the R225A and I343A mutants exhibit distinct, but divergent, changes to the conformational landscape. The R225A mutant increases the hydrophobicity of the interfacial site, stabilizing the *cis* conformation as the system maintains a ϕ value close to 0° for the entire simulation. Conversely, the I343A mutant reduces the hydrophobicity, lowering the energetic barrier between the *cis* and *trans* conformations, smoothing the distribution along ϕ. Ile52 is also an important hydrophobic residue in the MP–β-barrel connector portion of the interfacial site, and the I52A mutant shows a slight shift towards the transition state, suggesting a similar reduction in the energetic barrier as the I343A mutant. The K346A mutant is the stiffest of all the mutants. Lys346 resides within the hydrophobic cluster in the MP domain that holds its β-sheet structures together, and does not directly interact with the β-barrel domain. However, like the interfacial site between the MP and β-barrel domains, this lysine makes that hydrophobic cluster slightly amphiphilic. Mutating it to alanine stabilizes this cluster and stiffens the entire MP domain, particularly

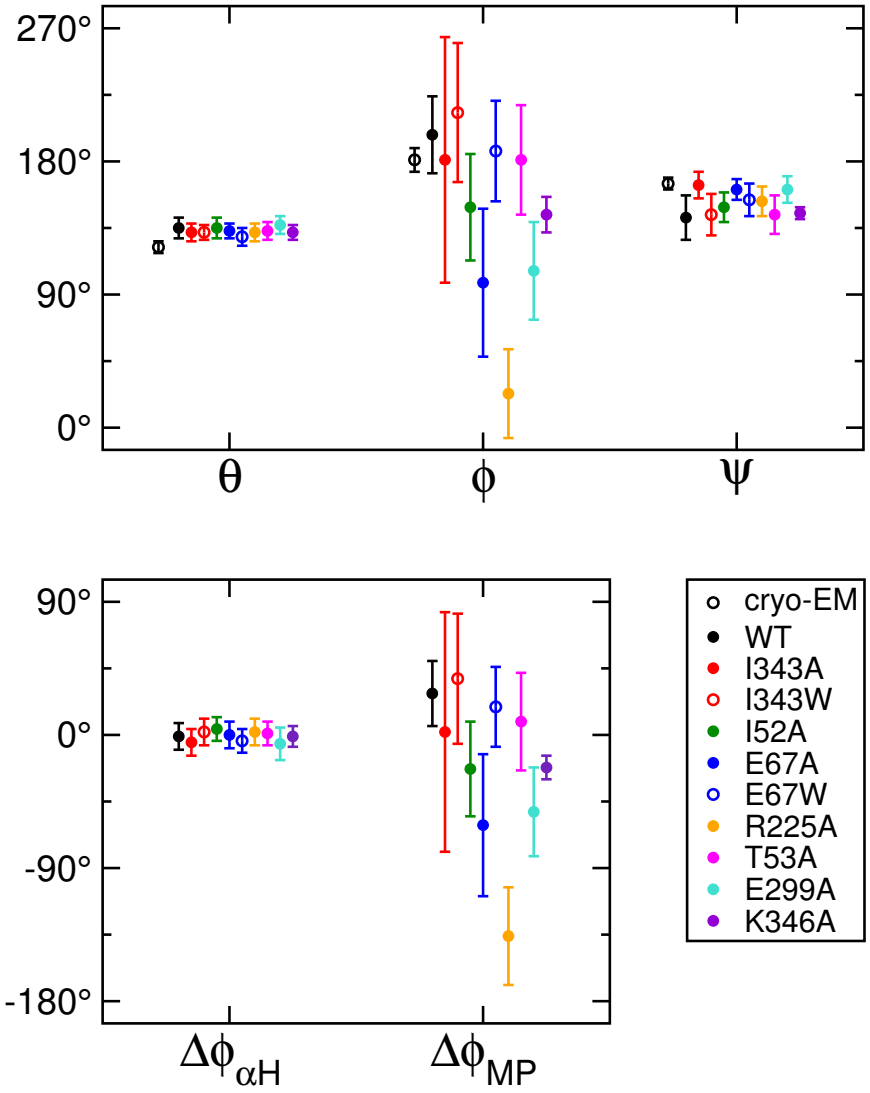


Figure 7.5: *Distribution of conformational angles from equilibrium simulations.* 250-ns equilibrium simulations of free AcrA monomer in water starting from a site-1 bound conformation from Ref [216]. Averages and standard deviations of θ, ϕ, and ψ were calculated from each simulation. For "cryo-EM", averages and standard deviations were calculated from all conformations of AcrA in Refs. [216] and [206]. $\Delta\phi_{\alpha\mathrm{H}}$ and $\Delta\phi_{\mathrm{MP}}$ are the changes in the α-hairpin and MP components, respectively, of the dihedral angle, ϕ, relative to the starting conformation. To calculate the $\Delta\phi$'s, we first aligned the β-barrel and lipoyl domains from every frame to the starting conformation, then calculated the change in the angles of the α-hairpin and MP domains along the β-barrel–lipoyl axis. It can clearly be seen that large changes in ϕ are almost entirely due to motions of the MP domain.

the connector strands.

7.3.5 2D PMFs of AcrA monomers bound to AcrB reveal a conformational selection mechanism for binding in one of the two binding sites

We also quantified the flexibility of AcrA after binding to AcrB. Starting from the AcrAB-TolC structure from Ref. [216], we built a system with the AcrB trimer embedded in a symmetric lipid bilayer representative of the *E. coli* inner membrane, and one copy of AcrA in either binding site 1 or 2. The AcrA MP domain was lipidated at its N-terminus, with the lipid tail also embedded in the bilayer (see Methods for more details). We first ran a 1-ns equilibrium simulation to get an initial set of windows to seed SLUS simulations. From those 1-ns simulations, we also extracted a set of contacts between AcrA and AcrB, defined as two residues that come within 5 Å. These contacts were restrained during the SLUS simulations to their average values in the equilibrium simulation in order to maintain AcrA in a bound state. These contacts restrict motion between the MP and β-barrel domains, particularly the ψ angle between the two domains. For this reason, our umbrella sampling simulations only bias the θ and ϕ angles, producing 2D PMFs.

Starting from an initial set of 6 windows for the site-1 system, we ran 25 iterations of the self-learning algorithm at 2 ns/window with a 3 kcal/mol cutoff, producing a total of 269 windows. The site-2 system produced a total of 411 windows after 47 iterations starting from an initial set of 28 windows. For both systems, we then ran 15 ns/window of REUS simulations, which were used to calculate the PMFs, for total simulation times of 4.6 μs and 7.0μs for sites 1 and 2, respectively. 2D PMFs for sites 1 and 2, as well as the θ-ϕ PMF of free AcrA, are plotted in Fig. 7.6. For both binding sites, we see a shift in the free energy minima from the free AcrA system towards the previously described transition region near $\phi = 150°$. There is also a distinct difference in the free landscapes between the two sites: site 2 is more flexible than site 1. While site 1 has one main conformational basin with a range of $\sim$90°-220° along ϕ (PMF $<$ 3 kcal/mol), site 2 has two large basins

that cover the entire range of ϕ. Site 1 reduces the conformational space of AcrA relative to the unbound state. This is indicative of conformational selection as the primary binding mechanism in this site. [221] Interestingly, the cryo-EM conformations are not located at the bound state minima. This could be due to either the lack of additional copies of AcrA (we only have one copy in our systems), or a lack of the outer-membrane channel, TolC.

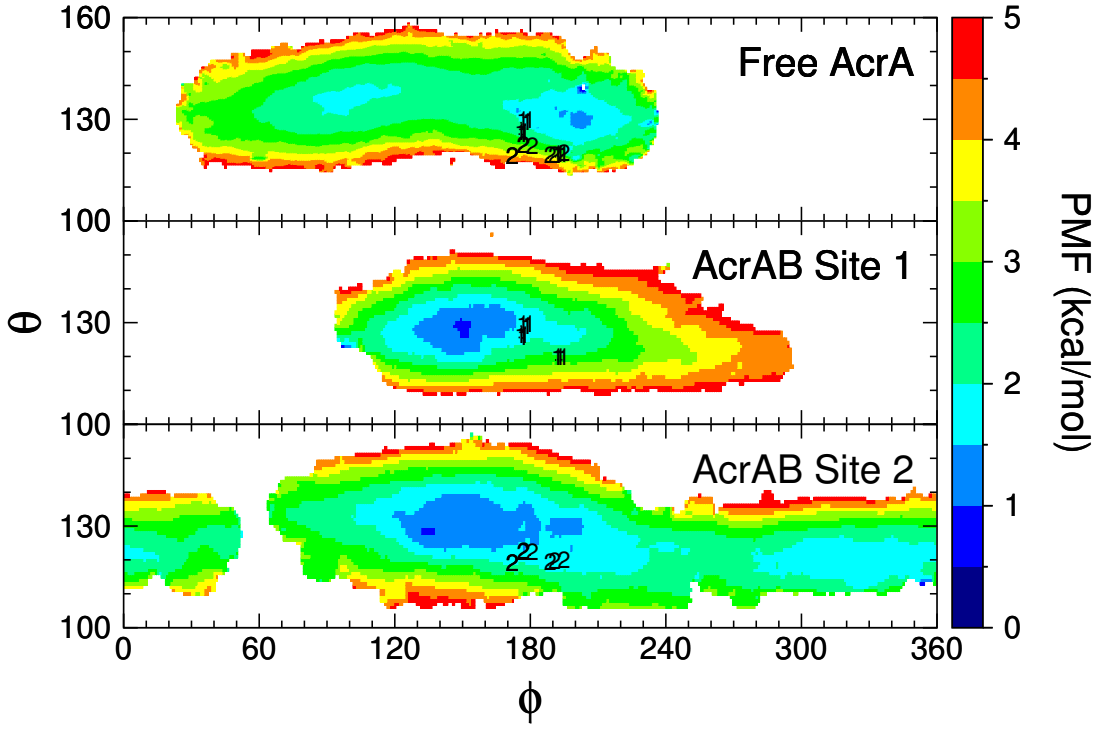


Figure 7.6: *2D PMFs of free and AcrB-bound AcrA*. The free AcrA PMF is a 2D projection of the full 3D PMF. Cryo-EM coordinates from Refs. [216] and [206] are also plotted with a "1" or a "2" for AcrA bound in site 1 or site 2, respectively.

Given their different conformational properties, the two copies within the AcrA dimer might play unique roles in the function of the efflux pump. The TriABC-OpmH triclosan efflux pump of *Pseudomonas aeruginosa*, for example, contains a unique heterodimeric membrane fusion protein, composed of TriA and TriB. Analogous mutations in the α-hairpin and MP domains on the two subunits were shown to have different effects on triclosan activity and pump assembly [222, 223]. Mutations in the α-hairpin domain in TriA, but not TriB, diminished pump activity. Conversely, mutations in the MP domain of TriB,

but not TriA, had a similar effect. Pump assembly and function is dependent on TriA's stabilizing interactions with the outer membrane channel and TriB's stimulation of the inner membrane transporter. The two copies of AcrA may be performing similar roles in the AcrAB-TolC efflux pump. Given that AcrA is stiffer in site 1 than site 2, site 1 could play the role of the channel stabilizer, with site 2 acting as the transporter stimulant.

7.4 Discussions

We have characterized the conformational landscape of AcrA, the membrane fusion protein (MFP) of the tripartite AcrAB-TolC multidrug efflux pump in *E. coli*, both in the free and AcrB-bound states. Free AcrA is very flexible, with its membrane proximal (MP) domain exhibit the widest range of motion of all its four structural domain. AcrA exists in two main conformations which are separated by a small energy barrier: a *trans* conformation where the MP and α-hairpin domains are pointing in opposite directions, defined by a dihedral angle of $150^\circ < \phi < 240^\circ$, and a *cis* conformation which encompasses all other ϕ angles. A set of amphiphilic contacts between the MP domain and the α-helix of the β-barrel domain, referred to as the interfacial site, define the *cis* conformation, whereas these contacts are broken in the *trans* conformation. In addition, the bound structures of AcrA as observed in cryo-EM structures of the fully assembled AcrAB-TolC complex are in the *trans* conformation and with almost no interfacial contacts. This suggests that this interface can be used as a switch to pick between binding compatible and incompatible states of AcrA. Our experiments show that mutations at the interfacial site which either stabilize interfacial contacts or lower the energy barrier for their formation increased the susceptibility of *E. coli* cells to the antibiotics to a similar degree as mutations at the hinge site between the hairpin and lipoyl domains, which is important for AcrA-TolC binding [16].

When bound to AcrB, the conformational landscape changes in a binding-site specific manner. While the second binding site exhibits a flexibility that is similar to that of the free AcrA monomer, the first site shows a sharp reduction in the accessible conforma-

tions, indicative of a conformational selection binding mechanism [221]. The difference in the conformational landscapes between the two binding sites also suggests that they two copies of AcrA play different roles in pump assembly and/or function. Efflux pumps with heterodimeric MFPs such as the TriABC-OpmH triclosan pump in *Pseudomonas aeruginosa* utilizes the two components at different steps of the assembly, with one responsible for stimulating transport activity at the inner membrane and the other for stabilizing pore formation at the outer membrane [222, 223]. The reduction in flexibility of AcrA in binding site 1 suggests its role may be for stabilizing the interactions with TolC, while the second site stimulates transport with AcrB.

It is important to keep in mind that for the full complex, six copies of AcrA must work in concert to perform efflux. For our work we have merely examined a single copy of AcrA at a time. While this may tell us some important information about initial binding events, more could be understood about pump formation from examining multiple copies, for example, 3 copies in each of the first or second binding sites on AcrB. In addition, how does AcrA effect the opening and closing of the OM pore? Comparison of closed and open conformations of TolC from the cryo-EM structures shows the disappearance of gaps between the copies of AcrA, through which antibiotics loaded into the barrel of the pump could leak out, when the pore is open [206]. While our work has shined light on possible binding mechanisms which occur early in the pump assembly process, it can be a guide for future computational and experimental studies on subsequent steps of the process.

CHAPTER 8
CONCLUSIONS AND FUTURE WORK

In this thesis I explored the folding and assembly of proteins within the cellular envelope of Gram-negative bacteria using molecular dynamics simulations and enhanced sampling techniques. To do this, I first extensively studied the folding processes of the small peptides, Ala_{10} and GB1, which are used to model α-helices and β-sheets, respectively. These model peptides were used to verify the most recent CHARMM protein force fields against substantial experimental data. Replica-exchange umbrella sampling (REUS) was used to calculate the folding free energies, and we observed that the CHARMM22* force field outperforms the CHARMM36, CHARMM36m, and CHARMM Drude polarizable force fields for β-sheets, while the Drude model performs the best for α-helical peptides. In addition, I examined the accuracy of different reaction coordinates commonly used in both simulation and experiments in reproducing the correct folding pathways. The common theme I observed is that measuring compaction of the peptide on its own is often insufficient, with some measure of structure acquisition, e.g. number the native hydrogen bonds, is required.

I then tackled the folding and assembly pathways of two large proteins in the cellular envelope, the pertactin autotransporter and the AcrAB-TolC multidrug efflux pump, so chosen for their capacity as new drug targets to combat antibiotic resistance. First, efflux pumps are a near-ubiquitous, primordial mechanism for antibiotic resistance in Gram-negative bacteria, and preventing assembly of these pumps, which span the cellular envelope, has lead to a new class of drugs, so-called efflux pump inhibitors (EPIs), which have the potential to revive the efficacy of old antibiotics in antibiotic resistance bacteria. Drugs binding to the membrane fusion protein (MFP), AcrA, which completes the assembly of the AcrAB-TolC efflux pump of bacteria by binding to both the inner membrane transporter, AcrB, and outer

membrane channel, TolC, where shown to significantly affect pump formation and increase the effectiveness of antibiotics. Molecular dynamics simulations and drug docking suggest that these drugs may be altering the flexibility of the protein to prevent binding to AcrB or TolC. Using 3D REUS simulations, I explored the conformational landscape of free AcrA monomers and observed two main conformations, a *trans*, bound-like conformation and a *cis*, non-bound conformation. Contacts between the two structural domains involved in binding to AcrB, called the interfacial site where some of the previously designed drugs were observed to bind, are responsible for the stabilization of the *cis* conformation, with these contacts vanishing in the *trans* conformation. However, due to the amphiphilic nature of these contacts, the stabilization is quite weak. Experimental mutational analysis of this site shows that altering the balance of these contacts can detrimentally affect pump formation, and simulations show that the conformational landscape is indeed shifted away from the bound state.

We also calculated the conformational landscape of AcrB-bound monomers of AcrA and observed distinct changes in the free energies. For both binding sites on AcrB, the free energy minimum was observed to be the transition state observed for the free AcrA monomer. However, the accessible conformations for the two sites were quite different. Site 1, which resides between two copies of AcrB, shows a reduced conformational space compared to the site 2, which interacts only with a single copy of AcrB. This suggests the two sites may have different functions in pump assembly and function, reminiscent of the heterodimer MFP, TriAB, from the TriABC-OpmH efflux pump, where TriA is involved in stabiizing the outer membrane channel, while TriB is used to stimulate the inner membrane transporter.

It is important to keep in mind that for the full complex, six copies of AcrA must work in concert to perform efflux. For this work I have merely examined a single copy of AcrA at a time. While this may tell us some important information about initial binding events, more could be understood about pump formation from examining multiple copies,

for example, 3 copies in each of the first or second binding sites on AcrB. In addition, how does AcrA effect the opening and closing of the OM pore? Comparison of closed and open conformations of TolC from the cryo-EM structures shows the disappearance of gaps between the copies of AcrA, through which antibiotics loaded into the barrel of the pump could leak out, when the pore is open [206]. While my work has shined light on possible binding mechanisms which occur early in the pump assembly process, it can be a guide for future computational and experimental studies on subsequent steps of the process.

Lastly, I have applied the results for the GB1 β-hairpin peptide to the β-helical passenger domain of the pertactin autotransporter. Autotransporters are virulence factors within infectious Gram-negative bacteria. The ubiquitous passenger domain β-helix has been implicated in the secretion process of virulence factors across the outer membrane into the extracellular space. Vectorial, C$\rightarrow$N folding of the helix is though to substantially increase the folding rate of the pertactin passenger domain, which in isolation folds much to slowly to be of use to the bacteria. I calculated the folding free energies of the different "bands" of the helix, which form the individual β-sheets of the helix, both in isolation and along the vectorial folding pathway, and found that the C$\rightarrow$N vectorial pathway significantly enhances both the stability of the bands and the rate of folding of the entire helix. Stability is achieved by two main factors: (1) decreasing the entropy of the unfolded state and (2) ensuring folded states stay close to the native state.

In this work, we only explored how whole, individual bands of the β-helix fold. However, this may not accurately represent the true folding pathway, either *in vitro* or *in vivo*. For example, perhaps only partial bands need to form for neighboring bands to begin folding on top of them. Multiple bands might also fold at the same time due to the collapse of a much larger hydrophobic core when the entire protein is present. In addition, we are most likely significantly underestimating the entropy of the unfolded states of the protein by excluding residues outside of the specific band being studied. Lastly, our chosen reaction coordinates, the number of native hydrogen bonds and the size of the hydrophobic

core, cannot account for the presence of non-native structures.

To further explore the folding pathways of the pertactin passenger domain, we will turn Markov state models (MSMs) amd machine learning (ML) algorithms. MSMs consist of running many long, unbiased simulations and calculating the number of transitions between clusters of states [224]. The advantage of using MSMs is that the reaction coordinates do not need to be explicitly defined *a priori*. Instead, they are learned from the simulation data, typically by calculating autocorrelation times. We will use a series of long unrestrained equilibrium simulations of larger constructs of the β-helix containing multiple bands.

Zimmerman and Bowman developed the fluctuation amplification of specific traits, or FAST, technique to explore the efficiently explore the conformational space, similar to SLUS, but without external biases [225]. Instead of guiding the system towards low energy paths, the FAST method attempts to maximize (or minimize) some variable, ϕ, while also encouraging exploration of lowly-sampled regions. (Some example variables for protein folding include the RMSD between the protein atoms and some reference structure, the number of protein-protein contacts, the number of water-protein contacts, or simply the total energy of the system). To do this, a MSM is first built from a collection of short simulations, and clustered into macrostates. Each macrostate is given a reward function:

$$r_\phi(i) = \bar{\phi}(i) + \alpha\bar{\psi}(i), \tag{8.1}$$

where $\bar{\phi}(i)$ is the normalized target value of macrostate i, such that it is 0 at its minimum (maximum) and 1 at its maximum (minimum); $\psi(i)$ is the occupancy of macrostate i; and $\bar{\psi}(i)$ is the normalized occupancy, with 0 being least occupied and 1 being most occupied. Typically, α is set to 1, although other systems might require lower values of α. A new set of simulations are spawned from the current set of macrostates, with the number of new simulations generated from each macrostate proportional to its reward, $r_\phi(i)$. A new MSM is built by combining the new and old simulations, and then the procedure is repeated until

the desired variable, ϕ, has been sufficiently sampled. The REAP method [**Shamsi**] builds upon the FAST method by allowing for multiple target variables and uses machine learning to decide which variable(s) to maximize (or minimize) at a given time.

We will use the time-structure based coordinate analysis (tICA) [226] to generate reaction coordinates for our system for clustering and free energy analysis. tICA has shown marked improvement over more traditional reaction coordinates [227]. Most recently, Ahalawat and Mondal have shown that using tICA to combine several traditional reaction coordinates for protein folding can more accurately describe the folding of the GB1 β-hairpin peptide than when using those reaction coordinates on their own by finding non-native kinetic traps [228]. First, a time-lagged covariance matrix,

$$\mathbf{C}(\Delta t) = \left\langle (\boldsymbol{\xi}(t) - \langle \boldsymbol{\xi}(t) \rangle)^t (\boldsymbol{\xi}(t + \Delta t) - \langle \boldsymbol{\xi}(t) \rangle) \right\rangle, \tag{8.2}$$

is calculated from the MD trajectories for a set of reaction coordinates, $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_N$, and used to solve the generalized eigenvalue equation,

$$\mathbf{C}(\Delta t)\mathbf{F} = \mathbf{C}(0)\mathbf{F}\mathbf{K}, \tag{8.3}$$

where $\mathbf{K} = \mathrm{diag}(k_1, k_2, \ldots, k_N$ and $\mathbf{F} = (\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_N$ are the eigenvalue and eigenvector matrices, respectively. The eigenvectors, or tIC's, with the largest eigenvalues are the slowest degrees of freedom in the system, and the two slowest tIC's will be used to generate free energy surfaces and clusters for our MSM as well as to guide our FAST/REAP procedure. We will also apply the variational Koopman operator method [229], which will allows to use shorter simulations to generate our tIC's.

FAST/REAP and tICA will not only allow us to further explore the folding pathways of pertactin, but it may also allow us to observe non-native structures that can act as kinetic traps, which also slow down the folding [227, 228]. We can also use these techniques to examine the role the β-domain has on passenger domain folding during secretion [197,

198]. We will build a model of the entire autotransporter with an OM-embedded β-domain and an unfolded passenger domain residing in the periplasm. The FAST/REAP method will be used to direct the passenger domain through the β-domain so that we can observe the early secretion process.

Appendices

APPENDIX A

SUPPLEMENTAL INFORMATION - THERMODYNAMICS OF DECA-ALANINE FOLDING IN WATER

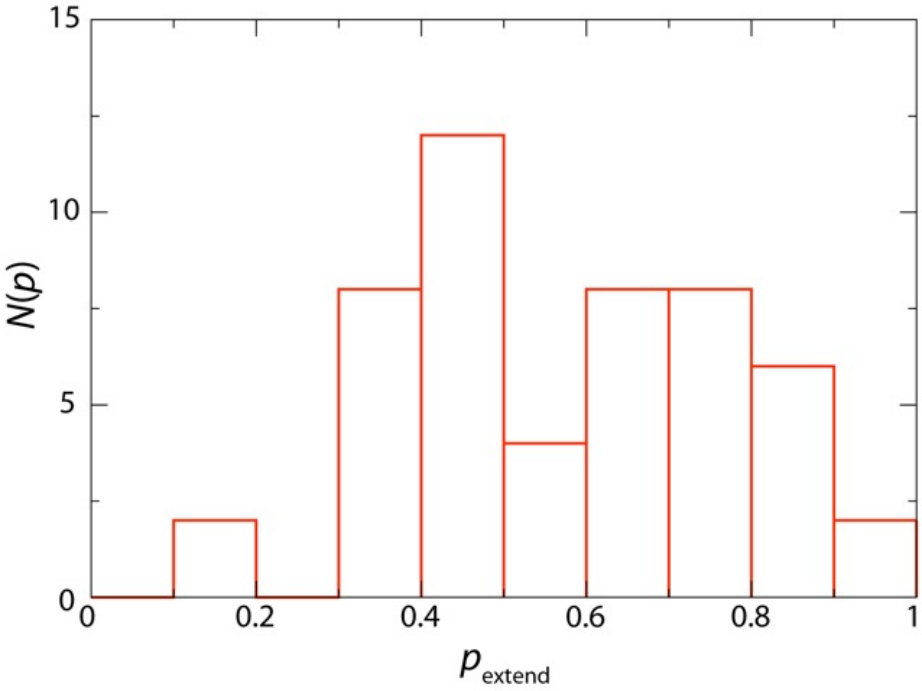


Figure A.1: Distribution of the committor function for 50 conformations near the free-energy maximum along the least free energy path. Each value of the committor was determined using 50 simulations of length 10 ps each.

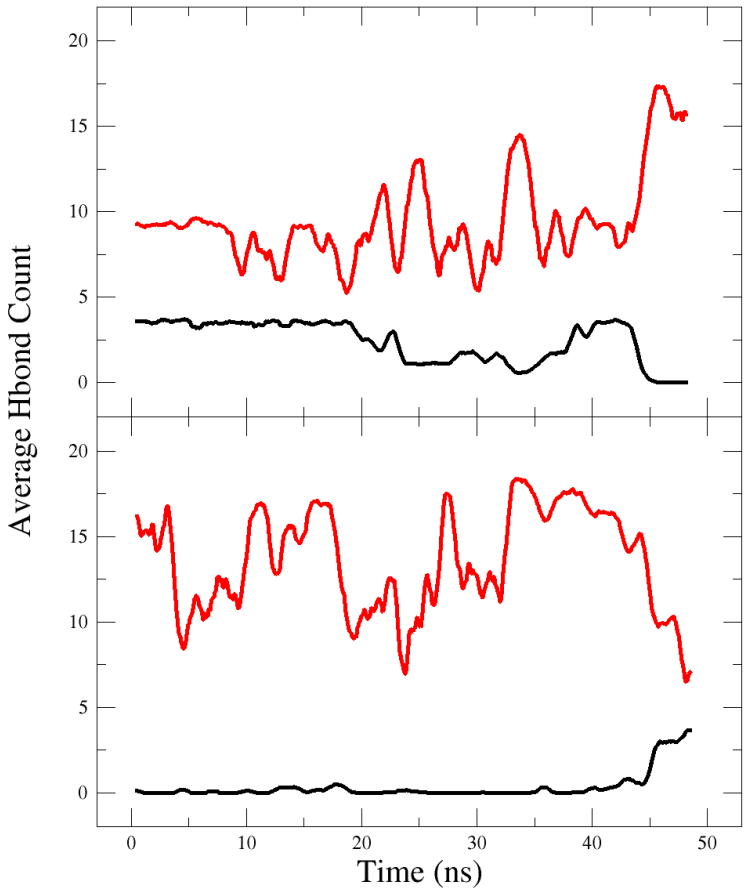


Figure A.2: Number of peptide-peptide (black lines) and peptide-water (red lines) hydrogen bonds during 50 ns equilibrium simulations of Ala_{10} in water. Top graph: α-helical starting state. Bottom graph: extended starting state.

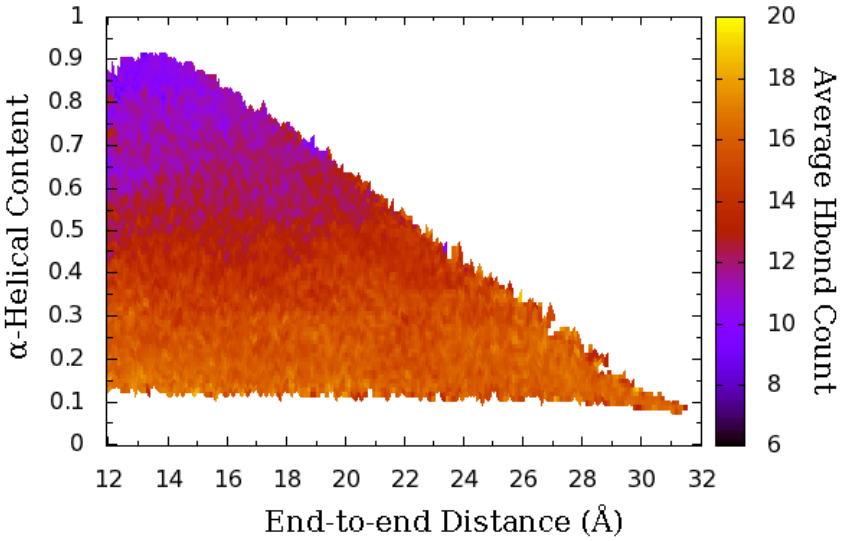


Figure A.3: Number of hydrogen bonds between the peptide backbone and water molecules for 2D REMD-US simulations of Ala_{10} in water.

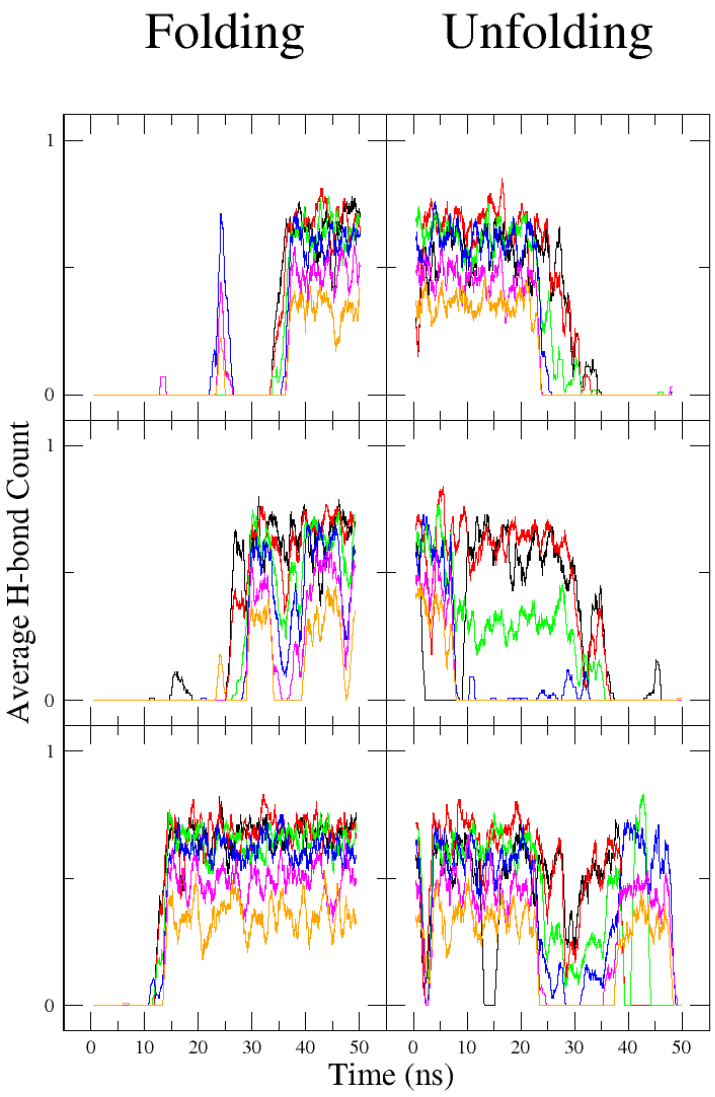


Figure A.4: Average number of $i, i+4$ hydrogen bonds for Ala_{10} during equilibrium folding and unfolding simulations. Tracking the average number of hydrogen bonds between residues 1 and 5 (black), 2 and 6 (red), 3 and 7 (green), 4 and 8 (blue), 5 and 9 (magenta), and 6 and 10 (orange) as a function of simulation time, with extended (left graphs) and α-helical (right graphs) starting states for a representative set of equilibrium simulations where folding or unfolding was observed, respectively.

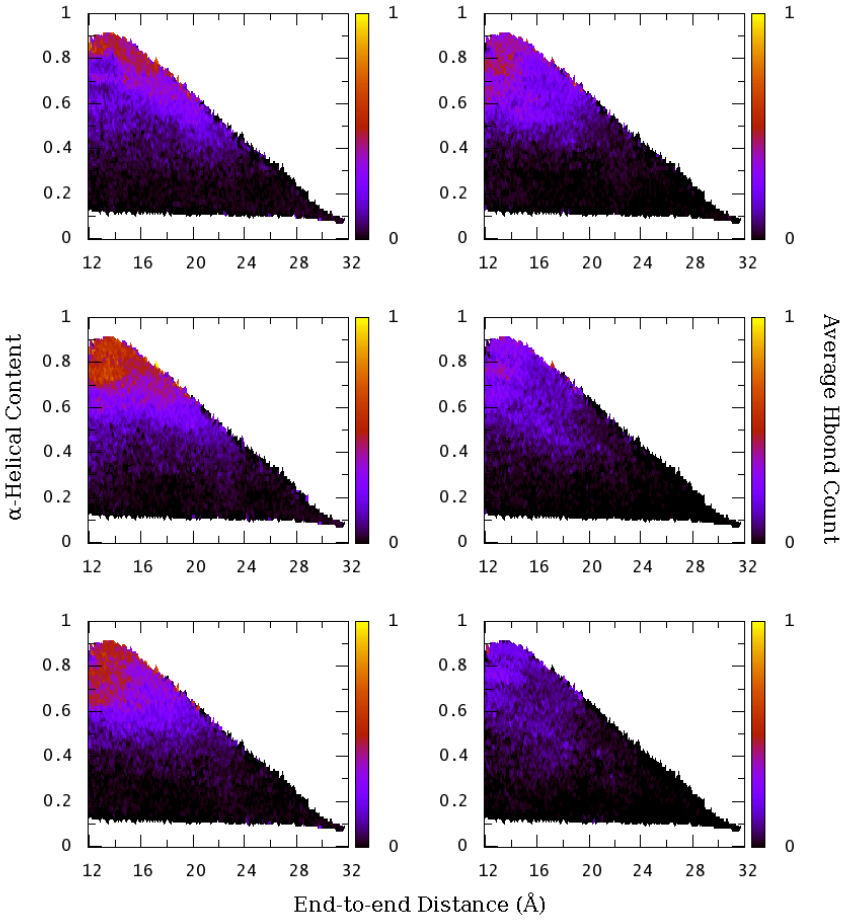


Figure A.5: Average number of $i, i+4$ hydrogen bonds for Ala_{10} during REMD-US simulations. Tracking the average number of hydrogen bonds between residues 1 and 5 (top left), 2 and 6 (top right), 3 and 7 (middle left), 4 and 8 (middle right), 5 and 9 (bottom left) and 6 and 10 (bottom right) as a function of end-to-end distance and α-helical content.

APPENDIX B

SUPPLEMENTAL INFORMATION - FOLDING FREE ENERGY LANDSCAPES OF β-SHEETS WITH NON-POLARIZABLE AND POLARIZABLE CHARMM FORCE FIELDS

Table B.1: Hydration energies of amino acid side chain models calculated using the C36 non-polarizable force field.

molecule	residue	ΔG_{elec}	ΔG_{disp}	ΔG_{rep}	ΔG_{total}	ΔG
n-butane	Ile	-0.02 ± 0.00	-11.00 ± 0.05	13.62 ± 0.12	2.60 ± 0.17	
isobutane	Leu	-0.07 ± 0.00	-10.80 ± 0.00	13.43 ± 0.05	2.56 ± 0.05	
methane	Ala	0.00 ± 0.00	-4.46 ± 0.00	6.84 ± 0.10	2.38 ± 0.10	
propane	Val	-0.05 ± 0.00	-8.98 ± 0.00	11.52 ± 0.07	2.49 ± 0.07	
acetamide	Asn	-9.06 ± 0.02	-9.47 ± 0.00	10.97 ± 0.02	-7.56 ± 0.05	$-$
p-cresol	Tyr	-6.12 ± 0.02	-15.75 ± 0.05	16.95 ± 0.14	-4.92 ± 0.22	$-$
ethanol	Thr	-6.93 ± 0.02	-8.01 ± 0.00	10.16 ± 0.05	-4.78 ± 0.07	$-$
methanethiol	Cys	-1.12 ± 0.00	-8.52 ± 0.00	9.37 ± 0.02	-0.27 ± 0.02	$-$
methanol	Ser	-6.86 ± 0.02	-5.94 ± 0.00	7.86 ± 0.05	-4.94 ± 0.07	$-$
methylethylsulfide	Met	-0.60 ± 0.00	-12.72 ± 0.02	13.91 ± 0.02	0.59 ± 0.05	$-$
3-methylindole	Trp	-5.59 ± 0.02	-19.08 ± 0.02	19.26 ± 0.05	-5.40 ± 0.10	$-$
methylimidazole	Hid	-11.59 ± 0.02	-11.97 ± 0.02	13.46 ± 0.14	-10.11 ± 0.19	-1
propionamide	Gln	-9.15 ± 0.02	-11.24 ± 0.00	12.81 ± 0.05	-7.59 ± 0.07	$-$
toluene	Phe	-1.74 ± 0.00	-14.57 ± 0.02	16.01 ± 0.14	-0.30 ± 0.17	$-$
acetate	Asp	-84.27 ± 0.05	-8.42 ± 0.00	10.35 ± 0.02	-82.35 ± 0.07	-8
n-butylammonium	Lys	-67.72 ± 0.07	-12.71 ± 0.02	14.44 ± 0.10	-66.00 ± 0.19	-6
n-propylguanidinium	Arg	-60.45 ± 0.02	-16.48 ± 0.02	16.73 ± 0.02	-60.21 ± 0.07	
propionate	Glu	-84.73 ± 0.02	-10.40 ± 0.00	12.38 ± 0.02	-82.75 ± 0.05	-7
n-methylacetamide	NMA	-8.41 ± 0.00	-12.36 ± 0.00	13.12 ± 0.07	-7.65 ± 0.07	-1

Table B.2: Hydration energies of amino acid side chain models calculated using the Drude polarizable force field.

molecule	residue	ΔG_{elec}	ΔG_{disp}	ΔG_{rep}	ΔG_{total}	ΔG
n-butane	Ile	-0.37 ± 0.01	-12.26 ± 0.01	15.32 ± 0.14	2.68 ± 0.15	
isobutane	Leu	-0.44 ± 0.00	-12.13 ± 0.03	15.07 ± 0.11	2.50 ± 0.14	
methane	Ala	-0.11 ± 0.00	-5.12 ± 0.00	7.46 ± 0.04	2.24 ± 0.05	
propane	Val	-0.33 ± 0.00	-9.95 ± 0.00	12.92 ± 0.14	2.65 ± 0.15	
acetamide	Asn	-12.08 ± 0.01	-9.89 ± 0.00	12.28 ± 0.10	-9.69 ± 0.11	$-$
p-cresol	Tyr	-6.11 ± 0.02	-18.15 ± 0.04	18.75 ± 0.08	-5.52 ± 0.14	$-$
ethanol	Thr	-6.59 ± 0.02	-8.75 ± 0.01	11.59 ± 0.05	-3.74 ± 0.08	$-$
methanethiol	Cys	-2.45 ± 0.01	-9.03 ± 0.02	10.44 ± 0.05	-1.04 ± 0.09	$-$
methanol	Ser	-5.79 ± 0.02	-6.50 ± 0.01	8.86 ± 0.08	-3.43 ± 0.11	$-$
methylethylsulfide	Met	-2.92 ± 0.02	-13.80 ± 0.03	15.60 ± 0.11	-1.12 ± 0.16	$-$
3-methylindole	Trp	-6.83 ± 0.04	-19.72 ± 0.02	21.44 ± 0.17	-5.11 ± 0.23	$-$
methylimidazole	Hid	-12.18 ± 0.04	-14.93 ± 0.03	15.45 ± 0.09	-11.67 ± 0.16	-1
propionamide	Gln	-10.76 ± 0.02	-12.09 ± 0.01	14.79 ± 0.07	-8.06 ± 0.11	$-$
toluene	Phe	-1.90 ± 0.01	-16.43 ± 0.01	17.85 ± 0.10	-0.48 ± 0.12	$-$
acetate	Asp	-85.45 ± 0.04	-12.30 ± 0.00	12.83 ± 0.06	-84.92 ± 0.11	-8
n-butylammonium	Lys	-63.31 ± 0.05	-14.78 ± 0.02	16.19 ± 0.08	-61.90 ± 0.16	-6
n-propylguanidinium	Arg	-58.93 ± 0.01	-16.52 ± 0.01	18.45 ± 0.05	-57.00 ± 0.07	
propionate	Glu	-81.53 ± 0.03	-14.40 ± 0.01	15.17 ± 0.14	-80.76 ± 0.19	-7
n-methylacetamide	NMA	-10.07 ± 0.03	-13.58 ± 0.01	15.15 ± 0.05	-8.50 ± 0.08	-1

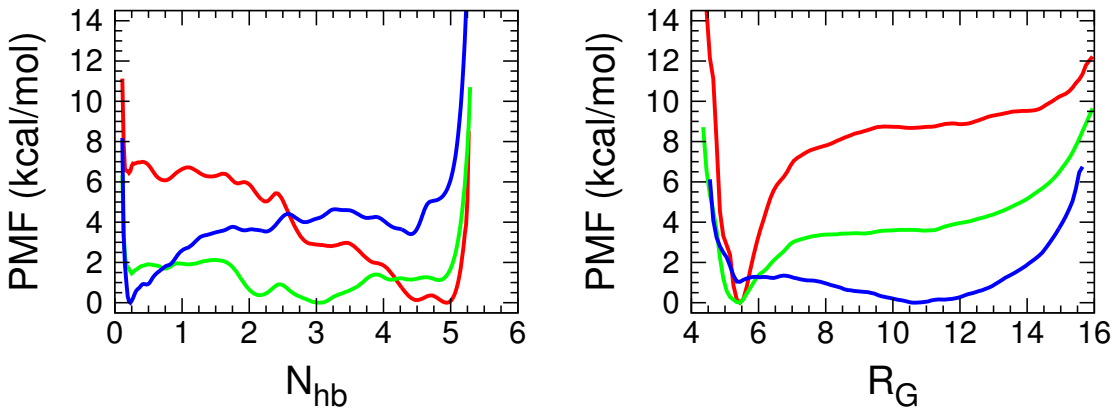


Figure B.1: *1D PMFs*. 2D PMFs were integrated along each reaction coordinate to produce 1D PMFs along the (left) N_{hb} and (right) R_{G} coordinates for the (red) C36, (green) C22*, and (blue) Drude-2013 force fields.

B.1 Adjusting Drude-2013 Partial Charges

Solubility in the Drude-2013 model is determined by both the intrinsic and the induced dipole moments. While the latter is determined by the polarizability and the charge of the Drude particle, the former is determined by the atomic partial charges. For the backbone amide and side chain hydroxyl groups, the C36 (and C22*) model has larger partial charges than the Drude-2013 model: $(q_{\mathrm{N}}^{\mathrm{Drude}}, q_{\mathrm{H}}^{\mathrm{Drude}}) = (-0.382e, 0.272e)$ versus $(q_{\mathrm{N}}^{\mathrm{C36}}, q_{\mathrm{H}}^{\mathrm{C36}}) = (-0.47e, 0.31e)$, and $(q_{\mathrm{O}}^{\mathrm{Drude}}, q_{\mathrm{H}}^{\mathrm{Drude}}) = (-0.46e, 0.36e)$ versus $(q_{\mathrm{O}}^{\mathrm{C36}}, q_{\mathrm{H}}^{\mathrm{C36}}) = (-0.66e, 0.43e)$. To test the effects of intrinsic dipoles on solubility and folding, we increased the Drude-2013 partial charges of these chemical groups to their C36 levels. When modifying atomic partial charges, we maintain the net charge of its chemical group. For example, for N-H groups, if we set $q_{\mathrm{N}}^{\mathrm{Drude}} \rightarrow q_{\mathrm{N}}^{\mathrm{C36}}$ (denoted $q_{\mathrm{N*H}}^{\mathrm{Drude}\rightarrow\mathrm{C36}}$), we recalculate $q_{\mathrm{H}}^{\mathrm{Drude}} = q_{\mathrm{N}}^{\mathrm{C36}} - q_{\mathrm{N\text{-}H}}^{\mathrm{Drude}}$; alternatively, if we set $q_{\mathrm{H}}^{\mathrm{Drude}} \rightarrow q_{\mathrm{H}}^{\mathrm{C36}}$ (denoted $q_{\mathrm{NH*}}^{\mathrm{Drude}\rightarrow\mathrm{C36}}$), we recalculate $q_{\mathrm{N}}^{\mathrm{Drude}} = q_{\mathrm{H}}^{\mathrm{C36}} - q_{\mathrm{N\text{-}H}}^{\mathrm{Drude}}$. Similar adjustments were made for the O-H groups. New hydration and folding free energies with these adjusted charges are shown in Fig. B.2.

Adjusting the atomic partial charge of the Drude-2013 amide H atom to its C36 value reproduced NMA hydration free energies fairly well, underestimating the free energy by

less than 0.5 kcal/mol; adjusting the Drude-2013 amide N atomic charge to its C36 value, however, overestimates the hydration free energy by more than 1 kcal/mol. The latter produces a larger intrinsic dipole than the former – $(q_{\mathrm{N}}, q_{\mathrm{H}}) = (-0.42e, 0.31e)$ for $q_{\mathrm{NH*}}^{\mathrm{Drude}\rightarrow\mathrm{C36}}$ versus $(-0.47e, 0.36e)$ for $q_{\mathrm{N*H}}^{\mathrm{Drude}\rightarrow\mathrm{C36}}$ – so the increase in intrinsic dipole moment going from $q_{\mathrm{N}}^{\mathrm{Drude}}$ to $q_{\mathrm{N}}^{\mathrm{C36}}$ is causing NMA to be too soluble. Similar results are observed for the O-H groups, where $q_{\mathrm{OH*}}^{\mathrm{Drude}\rightarrow\mathrm{C36}}$ overestimates of Ser and Thr side chain hydration free energies by 1 and 2 kcal/mol, respectively, but $q_{\mathrm{O*H}}^{\mathrm{Drude}\rightarrow\mathrm{C36}}$ overestimate the hydration free energies by almost 10 kcal/mol. The Drude-2013 GB1 folding free energy is reduced by $\sim$1 kcal/mol with C36 N-H charges, but increases significantly ($\sim$4-8 kcal/mol) for C36 O-H charges. For Ala_{10} the folding free energy is reduced by $\sim$1.5 kcal/mol for $q_{\mathrm{N*H}}^{\mathrm{Drude}\rightarrow\mathrm{C36}}$, but unaffected by $q_{\mathrm{NH*}}^{\mathrm{Drude}\rightarrow\mathrm{C36}}$. Increasing atomic partial charges appears to have a much greater effect on folding than increasing the atomic polarizabilities, although this depends on the system. For example, $q_{\mathrm{N*H}}^{\mathrm{Drude}\rightarrow\mathrm{C36}}$ partial charges and a 1.9$\times$ increase in N polarizability both yield $\sim$2 kcal/mol drop in NMA hydration free energies, but have opposite effects in the Ala_{10} and GB1 systems, with the folding free energies decreasing only with the former and latter parameters, respectively. For the O-H groups, while the GB1 folding free energy levels off at $\sim$3 kcal/mol after a 2$\times$ increase in the O atomic polarizability, it increases substantially when the intrinsic dipole moment is enhanced.

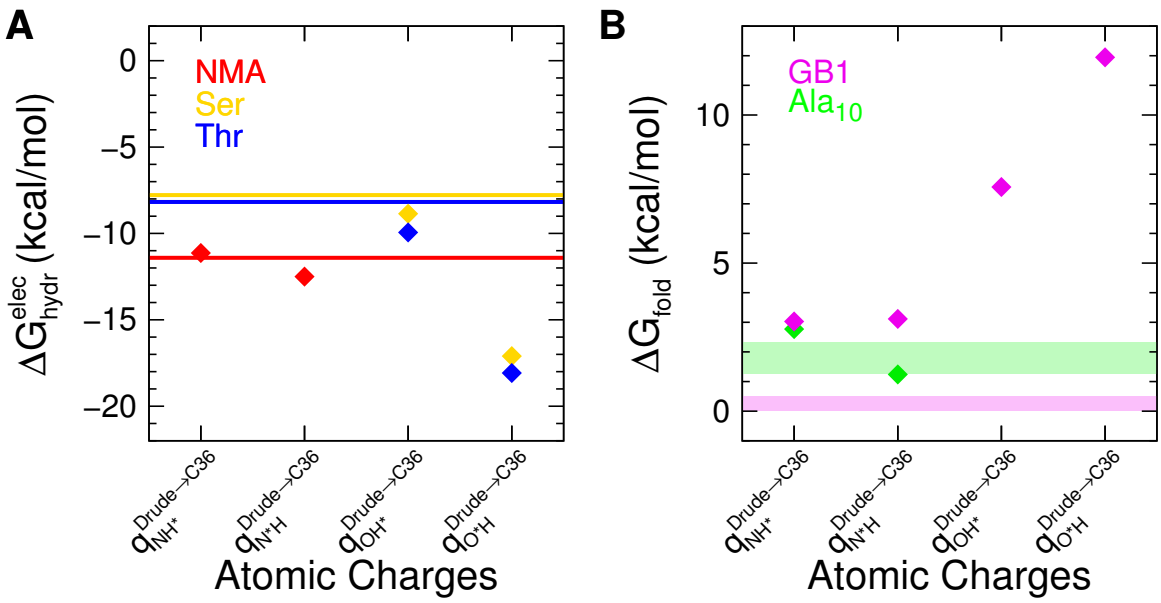


Figure B.2: *Hydration and Folding Free Energies for Adjusted Drude-2013 Atomic Partial Charges.* Atomic partial charges for backbone amide and side chain hydroxyl groups in the Drude-2013 model were adjusted to their C36 values. The partial charge of each atom was adjusted individually, denoted by (*), to its C36 value, while maintaining the net charge of its chemical group at its Drude-2013 value. (A) Electrostatic component of the hydration free energies for (red) NMA and (gold) Ser and (blue) Thr side chains with adjusted backbone amide (for NMA) or side chain hydroxyl (for Ser and Thr) partial charges. The solid lines show the target electrostatic component that would reproduce the experimental hydration free energies. (B) Folding free energies of (magenta) GB1 and (green) Ala_{10} calculated using reweighted REUS trajectories. The pink band gives the experimental folding free energies of GB1 from Refs. [122] and [130]. The light green band gives folding free energies of Ala_5 from Ref. [118], calculated by fitting MD simulations to experimental NMR J-coupling parameters.

B.2 Adjusting Drude-2013 Lennard-Jones Parameters

We also adjusted the Lennard-Jones (LJ) parameters for the backbone amide N and carboxyl O atoms. Chen and García showed that by decreasing the LJ minima between the RNA base pairs and the water molecules in the AMBER-99 force field by only 5%, they could prevent overstacking of the base pairs and reproduced RNA crystal structures with correct base pair hydrogen bonding [230]. We observe that making a similar, but opposite, adjustment between the backbone hydrogen bonding pairs – the amide N and carboxyl O atoms – and the SWM4-NDP water O atoms, we could decrease the folding free energy of GB1 towards its experimental values. Increasing the LJ minima by 4% (or 3% if accounting for the reduced data set bias), we could reproduce the experimental folding free energy of GB1 within about ± 1 kcal/mol, within the error of our free energy calculations. Since this would further underestimate the hydration free energy of NMA (see Table II in the main text), we also decreased the LJ minimum between the amide N atoms and the carboxyl O atoms. A decrease of the LJ minimum by 7% produces the experimental folding free energy of ~ 0 kcal/mol, not accounting for the reduced data set bias.

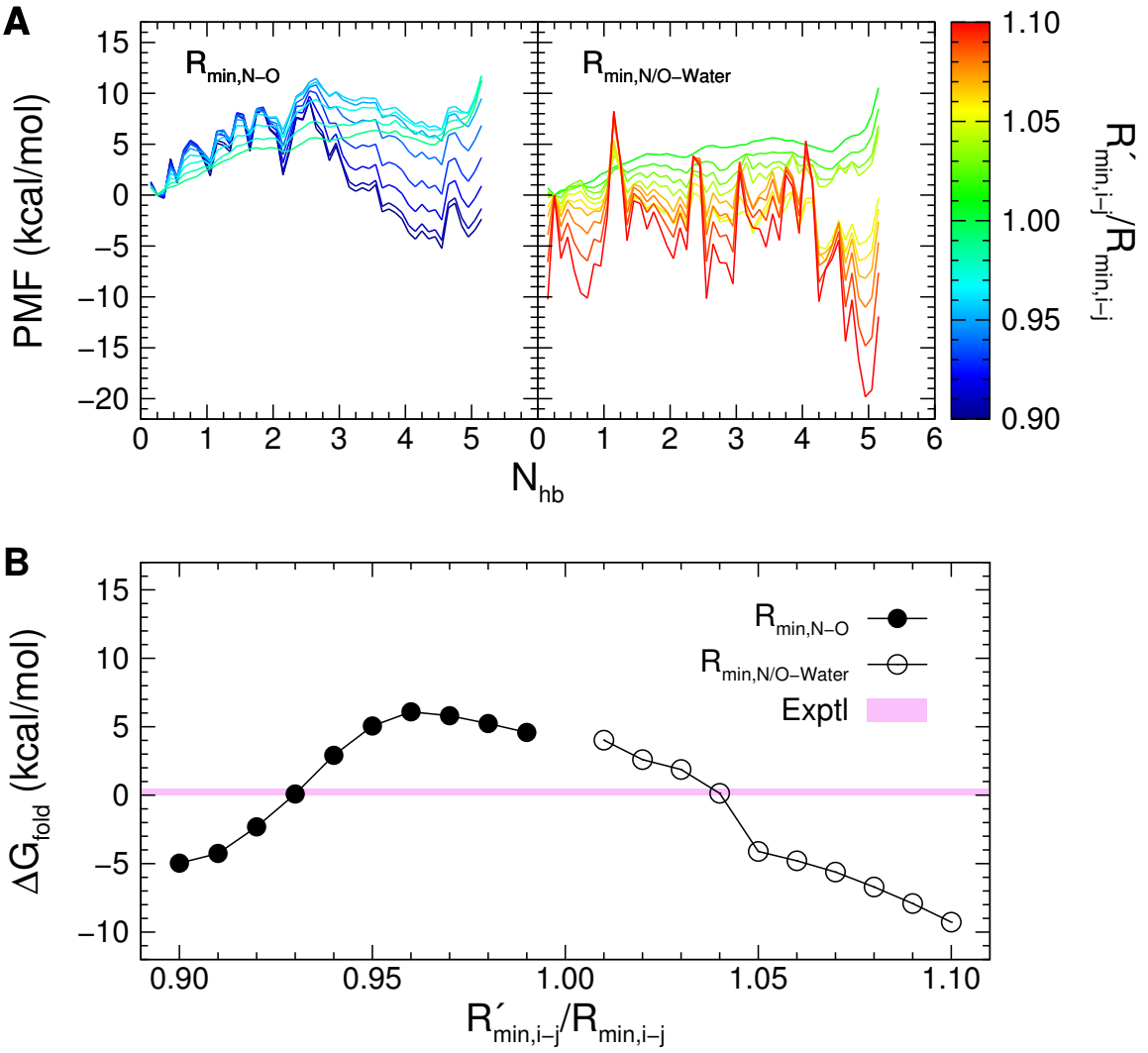


Figure B.3: *Free energies of GB1 Folding with adjusted Drude-2013 Lennard-Jones parameters.* (A) 1D PMFs of GB1 along the N_{hb} coordinate for altered amide N and carbonyl O Lennard-Jones parameters using 10-ps snapshots from last 5 ns/window of REUS simulations (see Methods). (Left) $R_{\mathrm{min,N\text{-}O}}$ reduced by up to 10%. (Right) $R_{\mathrm{min,N/O\text{-}Water}}$ increased by up to 10%. Lines are colored by ratio between new (R'_{min}) and old (R_{min}) parameters. (B) Folding free energies calculated from 1D PMFs in (A) (see Methods). (Black, solid circles) Adjusted $R_{\mathrm{min,N\text{-}O}}$. (Black, open circles) Adjusted $R_{\mathrm{min,N/O\text{-}Water}}$. (Pink band) Experimental folding free energies of GB1 from Refs. [122] and [130].

B.3 Dipole Moments

One advantage of the Drude polarizable force field is the reactivity of the molecular dipole moment to the surrounding electrostatic environment. Dipole moments for the Drude-2013 model were previously observed to be higher than those in the C36 model for both peptide backbones and side chains, particularly for GB1 backbone [33]. The increase in backbone dipole moments for the Drude model was shown to drive the cooperativity of helix formation in $(AAQAA)_3$, producing longer helices than for the C36 model at 300 K [143]. Given that the β-hairpin of GB1 is unstable in the Drude-2013 model yet is stable in the fixed-charge models, we measured the dipole moments of GB1 for our REUS trajectories to see how the dipole moment is playing a role in the stability of the hairpin.

The dipole moment of the backbone increases during folding for the Drude-2013 model, while it decreases for the non-polarizable models (see Fig. B.4). The inducible polarization of the Drude-2013 model mitigates the effect of the molecular geometry on the dipole moment. The total dipole moment, including the side chains, is generally higher in the folding region ($N_{\text{hb}} > 1$) for all three models, but Drude-2013 dipoles are significantly higher than those in the non-polarizable models, though they drop to C36 and C22* levels in the fully folded state ($N_{\text{hb}} = 5$). The sudden dip in dipole moment in the fully folded state could play a role in its lowered stability in Drude.

We studied the dipoles in more detail by examining the dipole moment of of the individual side chains (see Fig. B.5) for the three models. The most significant difference was observed in the Glu side chains, whose dipole moments are greater in the Drude-2013 model than in the non-polarizable models. This appears to be largely due to higher charges in the negatively charged groups of the side chains in the Drude model. For example, the carboxylate C atoms of Glu and Asp residues are more positive (0.70e) in the Drude-2013 model than in the C36 model (0.62e), while the O partial charges are the same in both models. The longer Glu side chain enhances the difference in dipole moments from

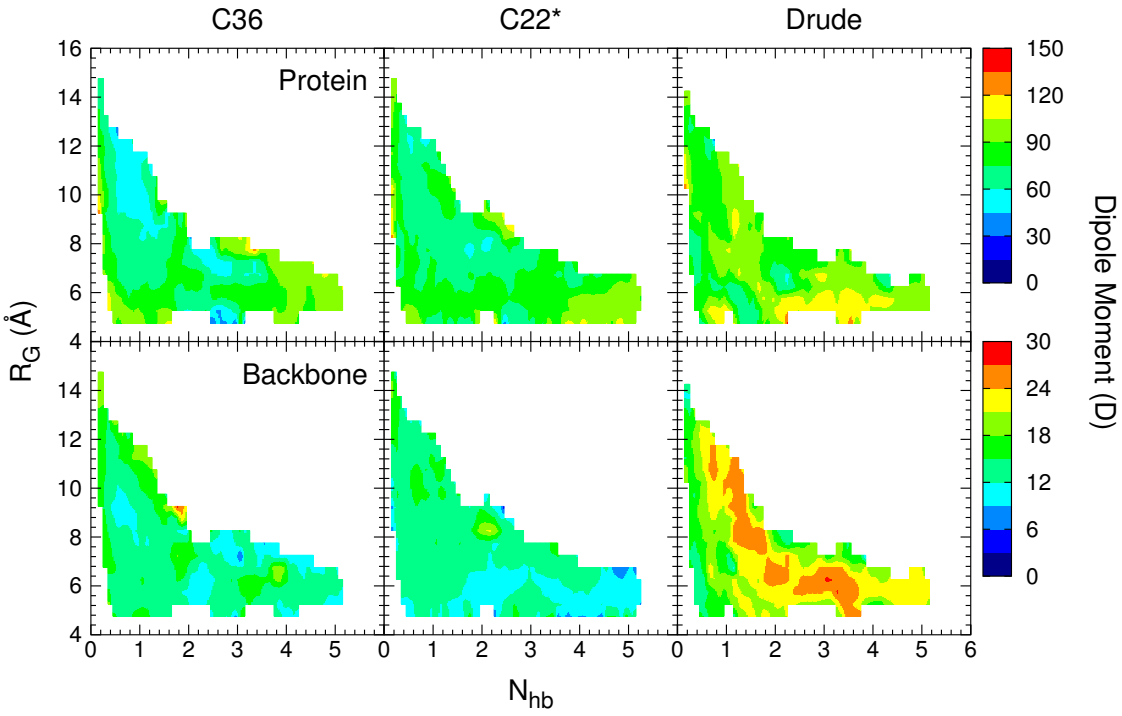


Figure B.4: *Dipole moments*. Dipole moments of GB1 calculated from REUS trajectories for C36 (left), C22* (center), and Drude (right). (Top) Dipole moment of entire protein. (Bottom) Dipole moment of backbone.

these redistributed charges, as the Drude-2013 Asp side chains are roughly the same as in the C36 model. Interestingly, while Glu and Asp side chain hydration energies are both overestimated by ∼2 kcal/mol for the C36 model, Drude-2013 Glu hydration free energies are accurately predicted, but Drude-2013 Asp hydration free energies are overestimated by 4 kcal/mol. Asp carboxylate partial charges are reduced in the C22* model compared to C36, while Glu partial charges are the same, and C22* produces better folding behavior than the C36 model. A similar adjustment of Asp carboxylate partial charges in the Drude-2013 model could also improve its folding behavior.

We also examined backbone N-H and C=O dipole moments in the Drude-2013 model to see the extent of hydrogen bond polarization during folding (see Fig. B.6A). We separated out the native hydrogen-bonding residues from the non-hydrogen-bonding residues. For the non-polarizable models, N-H and C=O bond dipole moments are fixed at values of 1.88 D and 3.02 D, respectively, with larger C and O partial charges leading to larger

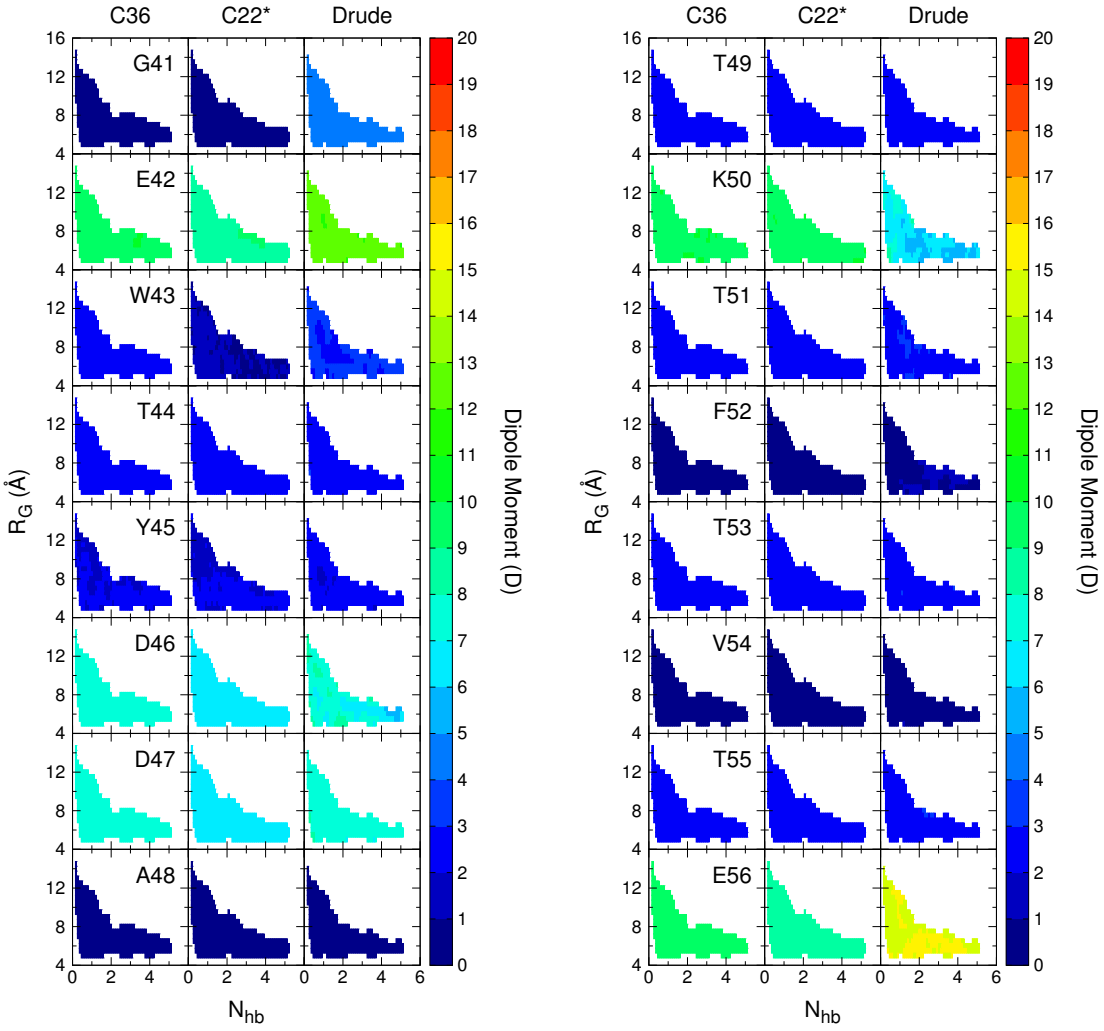


Figure B.5: *Dipole moments for each side chain of GB1.* Dipole moments of each side chain of GB1 calculated from REUS trajectories for (left) C36, (center) C22*, and (right) Drude-2013 models.

dipole moments in C=O bonds than N-H bonds. For the Drude-2013 model, N-H bond dipole moments range from 1.89-2.27 D, while C=O bond dipole moments range from 3.18-3.39 D. Native hydrogen-bonding N-H bonds exhibit larger dipole moments in the folded state than other residues. For C=O bonds, non-hydrogen-bonding residues actually exhibit larger dipole moments, with native hydrogen-bonding C=O dipole moments increasing to a much lesser extent in the folded state than their N-H counterparts. Another advantage of the Drude-2013 model is that the direction of the dipole moment can change as well as its magnitude. We measured the deviation of the dipole moment vectors from the bond vectors and found that N-H bond dipole moments deviate more from the bond direction ($\sim 13^{\circ}$) than C=O bond dipoles ($\sim 4^{\circ}$) (see Fig. B.6B). We then looked at the individual atomic dipole moments for the N-H and C=O atoms. As expected from the N-H and C=O bond dipole moments, N atoms show larger increases in the magnitudes of their dipole moments during folding for native hydrogen-bonding residues, while C and O atomic dipole moments are higher in non-hydrogen-bonding residues. For all native hydrogen-bonding residues, the atomic dipole moments become more aligned with bond vectors during folding, whereas non-hydrogen-bonding deviations remain relatively unchanged, with the exception of C atoms, which actually deviate even more during folding.

It's clear that polarization of the native hydrogen bonds during folding is driven by a combination of increasing atomic polarization and greater alignment of the atomic dipoles. The more interesting result is why C=O dipole moments are higher in non-hydrogen-bonding residues. This could be the result of water–peptide hydrogen bonding. SWM4-NDP dipole moments are 2.46 D [231]. This is significantly higher than the N-H dipole moment. Water–peptide hydrogen bonding is also much higher in the Drude model than in the non-polarizable models. The increase in C=O dipole moments for the non-hydrogen-bonding residues, and a lack of similar increase in the N-H dipole moments for these same residues, could be the result of much stronger hydrogen bonding of the water with backbone C=O groups than N-H groups. A stronger dipole in SWM4-NDP water molecules

than in TIP3P water molecules (2.347 D, see Ref. [92]) combined with polarizable C=O bonds could make C=O–water hydrogen bonding much more favorable in the Drude-2013 model than the non-polarizable models. N-H polarization energies need to be comparable to C=O polarization energies for C=O bonds to not dominate the interactions. This could be accomplished by increasing the extent of induced polarization in the N-H bonds in several possible ways: (1) increasing the N and H partial charges, (2) increasing the N polarizabilities, or (3) moving the hard wall potential between N parent and Drude atoms further away. We have explored the first option in Section A above and the second option in Section IIIC in the main text.

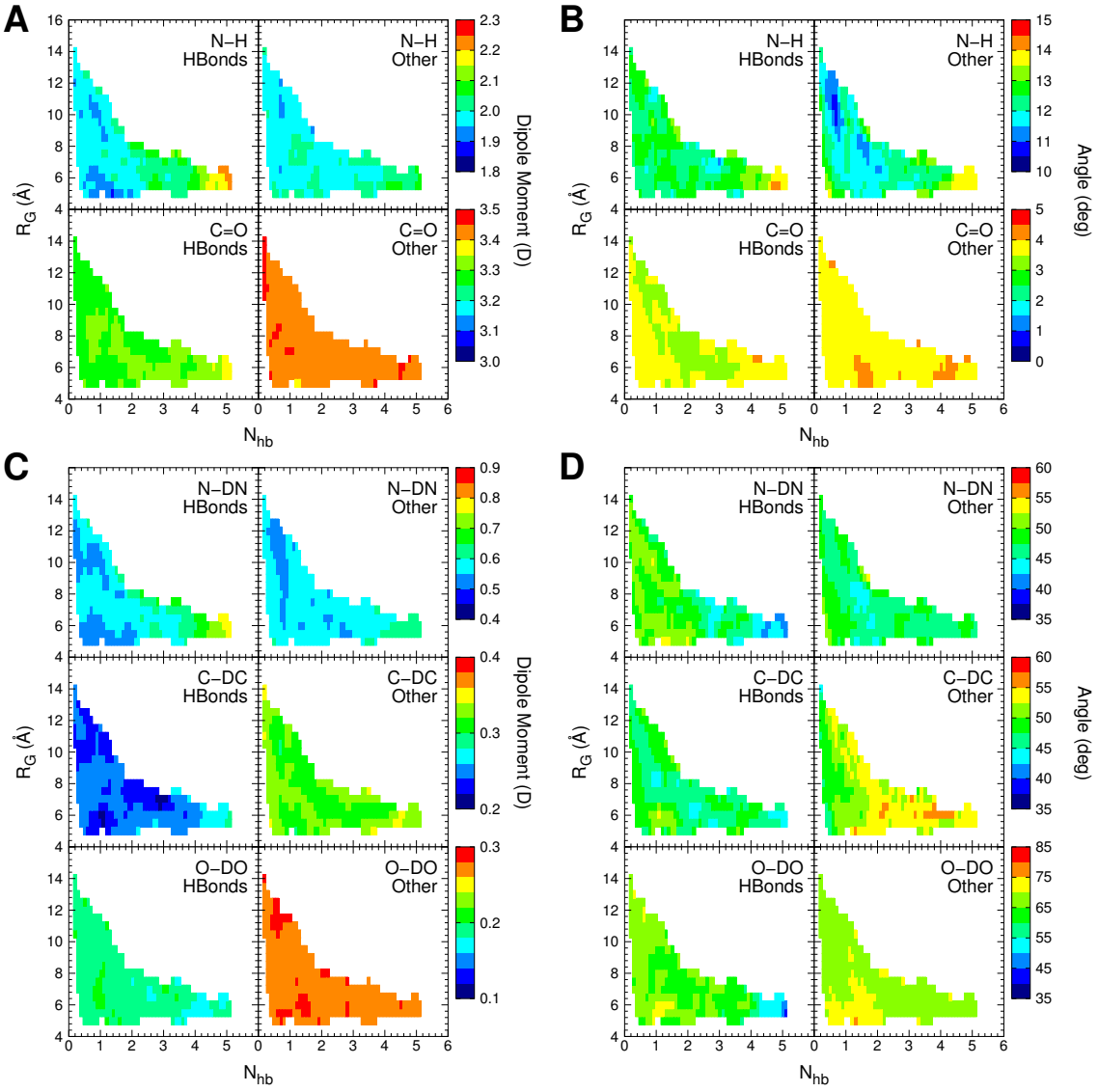


Figure B.6: *Dipole moments of Drude backbone N-H and C=O bonds*. (A) Magnitude of (top) N-H and (bottom) C=O dipole moments for (left) native hydrogen-bonding residues (Glu42, Thr44, Asp46, Thr51, Thr53, and Thr55) and (right) all other residues. For C36 and C22*, N-H and C=O bonds have dipole moments of ∼1.88 D and ∼3.02 D, respectively. (B) Angle of dipole moment relative to the bond vector (N→H for N-H bonds and (O→C for C=O bonds). (C) Magnitude of atomic dipole from parent atom and Drude particle for N, C, and O atoms. (D) Angle of atomic dipole moment relative to bond vector.

B.4 Ala_{10} Folding Free Energies

All-atom simulations of non-polarizable Ala_{10} were performed starting from the 104-atom compact helical model used by Park et al [86], capped with an acetylated N-terminus and amidated C-terminus. For non-polarizable models, the visualization and analysis program VMD [91] was used to place Ala_{10} in a water box of 10,850 TIP3P (Ref. [92]) water molecules with dimensions $70 \times 70 \times 70$ Å^3, resulting in a system of 32,659 atoms. Molecular dynamics simulations were carried out using NAMD 2.10-12 (Ref. [**Phillips2005**]) with CHARMM all-atom force fields C22* (Ref. [144]) and C36 (Refs. [94, 95]). Simulation parameters were identical to those used for GB1.

For the polarizable model, a non-polarizable model was first built using CHARMM-GUI [149]. Similar to the previous non-polarizable models, Ala_{10} was solvated in a water box of dimensions $70 \times 70 \times 70$ Å^3 with 10,798 water molecules. The Drude Prepper [149] from CHARMM-GUI was then used to convert the non-polarizable model into the Drude-2013 polarizable model for the protein [33] solvated in SWM4-NDP polarizable water molecules [151], resulting in a system of 54,175 atoms, including Drude particles. We used the Drude-2013 parameters released in July 2015. The system was then minimized and pre-equilibrated using the NAMD input scripts provided by Drude Prepper. Simulation parameters were identical to those used for GB1.

Two-dimensional potentials of mean force (PMF) were calculated using umbrella sampling with replica-exchange molecular dynamics (REUS) [51]. Reaction coordinates were calculated and biased using the *collective variables* (colvars) module of NAMD 2.10 (Ref. [154]). The first reaction coordinate is the α-helical content calculated using the `alpha` colvar:

$$\alpha(\mathrm{C}_\alpha^{(N_0)}, \mathrm{O}^{(N_0)}, \mathrm{C}_\alpha^{(N_0+1)}, \mathrm{O}^{(N_0+1)}, \mathrm{N}^{(N_0+5)}, \mathrm{C}_\alpha^{(N_0+5)}, \mathrm{O}^{(N_0+5)}, \ldots, \mathrm{N}^{(N_0+N)}, \mathrm{C}_\alpha^{(N_0+N)} = \frac{1}{2(N-2)} \sum_{n=N_0}^{N_0+N-2} \mathrm{angf}\big(\mathrm{C}_\alpha^{(n)}, \mathrm{C}_\alpha^{(n+1)}, \mathrm{C}_\alpha^{(n+2)}\big) + \frac{1}{2(N-4)} \sum_{n=N_0}^{N_0+N-4} \mathrm{hbf}\big(\mathrm{O}^{(n)}, \mathrm{N}^{(n+4)}\big), \tag{B.1}$$

where N is the number of helical residues, N_0 is the index of the first residue, hbf is the hydrogen bond scoring function given by Eq. 1 in the main text, and angf is the angle scoring function given by

$$\text{angf}\big(\text{C}_\alpha^{(n)}, \text{C}_\alpha^{(n+1)}, \text{C}_\alpha^{(n+2)}\big) = \frac{1 - \big(\theta(\text{C}_\alpha^{(n)}, \text{C}_\alpha^{(n+1)}, \text{C}_\alpha^{(n+2)}) - \theta_0\big)^2/(\Delta\theta_{\text{tol}})^2}{1 - \big(\theta(\text{C}_\alpha^{(n)}, \text{C}_\alpha^{(n+1)}, \text{C}_\alpha^{(n+2)}) - \theta_0\big)^4/(\Delta\theta_{\text{tol}})^4}, \tag{B.2}$$

where $\theta_0 = 88^\circ$ is the reference α-helical angle, and $\Delta\theta_{\text{tol}} = 15^\circ$ is the angle cutoff relative to the reference angle. $\alpha \in [0, 1]$, with $\alpha = 0$ indicating no helical content within the N residues, and $\alpha = 1$ indicating all N residues are perfectly helical. The second reaction coordinate is the end-to-end distance between the first and last backbone carbonyl C atoms. 204 windows were simulated for 12 ns/window (Drude-2013) and 20 ns/window (C36 and C22*), for a total simulation time of 2.4-4.1 μs per system. Windows were spaced along the α coordinate by 0.05 using a 500 kcal/mol harmonic force constant. Windows were spaced along the end-to-end distance coordinate by 1.0 Å, using a 5 kcal/mol·Å^2 harmonic force constant. The first 2 ns/window of the REUS simulations were omitted when calculating the PMF, which was generated via the weighted histogram analysis method (WHAM) [155]. Starting states for each window were generated using steered molecular dynamics (SMD) [156]. Our previous work demonstrated that the α coordinate is a better measure of helix folding for Ala_{10} than the end-to-end distance coordinate [34], so folding free energies are calculated along the α coordinate unless otherwise indicated.

Table B.3: ΔG_{fold} for Ala_{10} calculated from 1D PMFs in kcal/mol.

Coordinate	C36	C22*	Drude
α-Helical content	$+0.86 \pm 0.66$	$+0.63 \pm 0.32$	$+2.40 \pm 0.88$
End-to-end distance	$+0.98 \pm 0.90$	-0.85 ± 3.41	$+2.30 \pm 0.79$

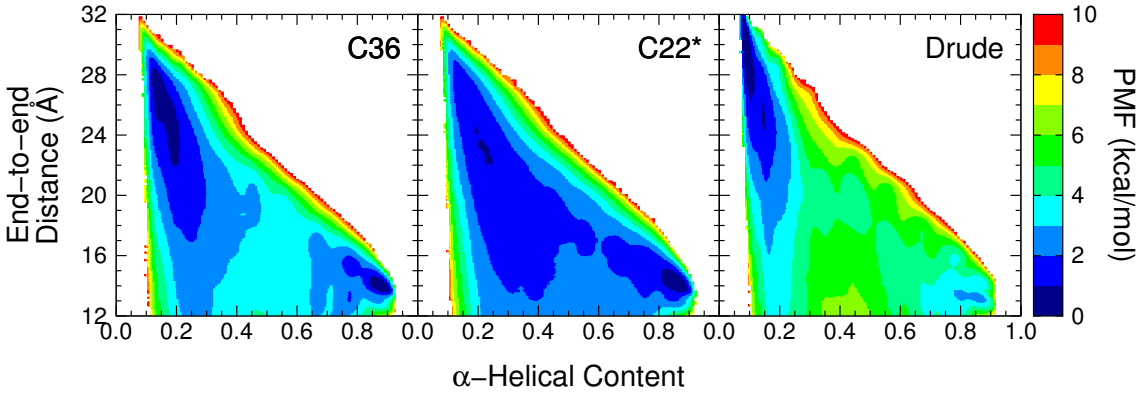


Figure B.7: *Folding free energy landscapes of Ala_{10}.* 2D PMFs calculated by REUS using (left) C36, (middle) C22*, and (right) Drude-2013. A total of 204 windows were utilized and simulated for 12 ns/window and 20 ns/window for the Drude-2013 and C36/C22* systems, respectively. The first 2 ns/window of each system was omitted when calculating the PMFs.

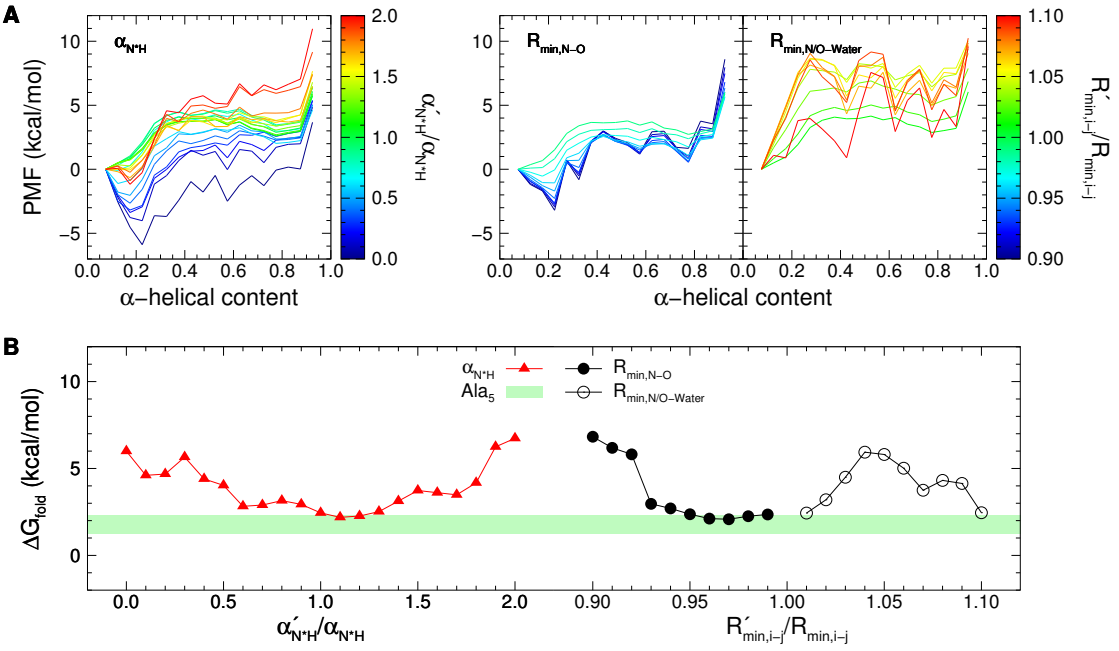


Figure B.8: *Free energies of Ala_{10} folding with adjusted Drude-2013 polarizabilities and Lennard-Jones parameters.* (A) 1D PMFs of Ala_{10} along the α coordinate for adjusted Drude-2013 (left graph) backbone amide N-H polarizabilities and partial charges and (right two graphs) backbone amide N and carbonyl O Lennard-Jones parameters using 10-ps snapshots from last 10 ns/window of REUS simulations (see Methods). For original Drude-2013 parameters ($\alpha'/\alpha = 1$ in the left graph), the PMF was also recalculated using this reduced data set. (Solid lines) 1D PMFs for adjusted parameters, colored by ratio between new (α' and R'_{min}) and old (α and R_{min}) parameters. (Black dashed line, left graph) 1D PMF for C36 H atomic charges, $q^{\text{C36}}_{\text{NH*}}$. (Black dotted lines, left graph) 1D PMF for C36 N atomic charges, $q^{\text{C36}}_{\text{N*H}}$. (B) Folding free energies calculated from 1D PMFs in (A) (see Methods). (Red, upward-facing triangles) amide N polarizabilities. (Black, solid circles) Adjusted $R_{\text{min,N-O}}$. (Black, open circles) Adjusted $R_{\text{min,N/O-Water}}$. (Light green band) Folding free energies of Ala_5 from Ref. [118], calculated by fitting MD simulations to experimental NMR J-coupling parameters.

B.5 Hydrogen Bonding and Cooperativity

One important factor governing folding free energies is the folding pathway itself. The presence of water molecules, for example, can alter the folding pathway of helices from their vacuum folding pathway [34, 109]. We examined the ordering of both peptide–peptide and peptide–water hydrogen bonding during GB1 β-hairpin formation for our C36, C22*, and Drude-2013 REUS simulations (Fig. B.9). There is a stark increase in backbone–water hydrogen bonding in the Drude-2013 model compared to both the C36 and C22* models (Fig. B.9, top), with roughly 10-15 more hydrogen bonds occupied at any one time in the Drude-2013 model than in the fixed-charge models. Splitting of the backbone carbonyl O charge into two lone pairs may allow for more optimal hydrogen bonding with multiple water molecules. In addition, backbone carbonyl O and SWM4-NDP H atoms have larger partial charges than in the fixed-charge force fields, and the Lennard-Jones minimum for SWM4-NDP is lower in energy than that of TIP3P (Ref. [151]). The overall net negative charge of the C=O group, compared to a neutral C=O group in the fixed-charge models, could also be enhancing electrostatic interactions. Atomic polarization will also contribute stronger dipole-dipole interactions. Khan et al. had previously shown that some charged and polar moieties in Drude-2013 have a tendency to be too close together, with interaction energy minima shifted towards smaller separations compared with QM calculations and a modified version the C36 force field [232]. Interactions between polarizable water molecules and a polarizable backbone could be producing a similar phenomenon for the GB1 peptide, resulting in higher overall peptide–water hydrogen bonding.

Next, we examined the intrapeptide hydrogen bond ordering (Fig. B.9, bottom). The general trend appears to be the same for all three force fields. One of the inner most hydrogen bonds (Asp46–Thr51) forms first, followed by the central hydrogen bonds (Thr44–Thr53), and the outermost hydrogen bonds (Glu42–Thr55) forming last, or not at all, in the fully folded native state. There are some subtle differences, however, in the bond order-

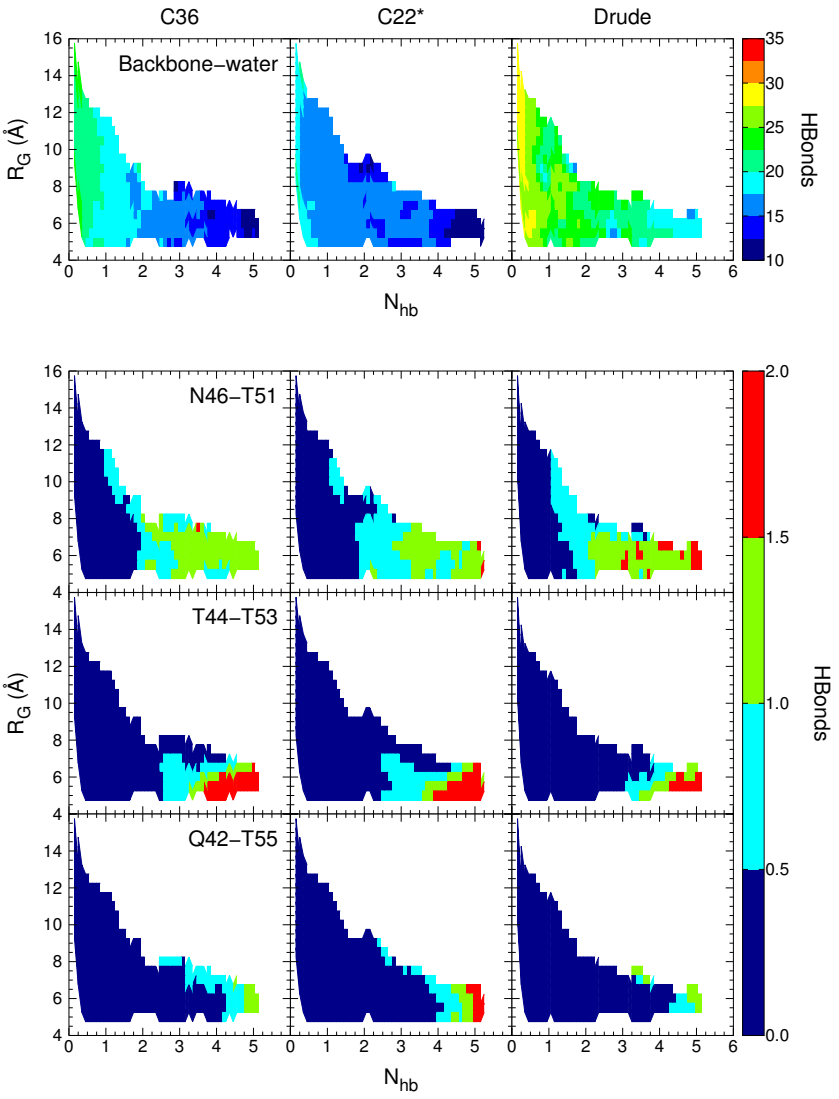


Figure B.9: *Hydrogen bonding stucture of GB1.* Number of hydrogen bonds calculated from REUS trajectories for C36 (right), C22* (center), and Drude (left). (Top group) Total number of hydrogen bonds between the protein backbone and surrounding water molecules. (Bottom group) Number of hydrogen bonds between residues (top) Asp46 and Thr51, (middle) Thr44 and Thr53, and (bottom) Glu42 and Thr55. Hydrogen bonds are defined to be occupied if the N–O distance is less than 3.4 Å and the N-H–O angles is less than 30°.

ing. While the non-polarizable models have nearly identical hydrogen bond ordering, an additional inner hydrogen bond forms in the Drude-2013 model that is not formed in the non-polarizable models. As a consequence, the central hydrogen bonds form slightly later in the folding process. Even though bond ordering is mostly conserved, folding does not appear to be cooperative in the Drude model; formation of the first hydrogen bond does not lead to a downhill folding event as is observed in the non-polarizable models. Cooperativity was previously shown to be improved in the Drude model for the α-helical peptide, $(AAQAA)_3$, because of enhanced backbone dipole moments over the C36 model, resulting in longer, more stable helices at 300 K [143]. The enhancement to the dipole moment from hydrogen bonding for residues in β-sheet conformations, however, was not only much less than that from residues in α-helical conformations, but was even slightly less than the enhancement from simply being in an extended conformation, like a PPII helix (0.76 D, 0.27 D, and 0.21 D for α-helical, PPII, and β-sheet conformations, respectively) [233]. In α-helical conformations, all hydrogen bonding groups align parallel to each other and the backbone, enhancing the dipole moment of the backbone; however, in β-sheet conformations, hydrogen bonds are anti-parallel within the same residue, and align perpendicular to the backbone, so induced polarization of the N-H and C=O groups during intrapeptide hydrogen bonding does not add to the dipole moment of the backbone to the same extent as it does for α-helices. Therefore, cooperative folding for β-sheets is not driven by enhanced backbone dipoles as it is for α-helices.

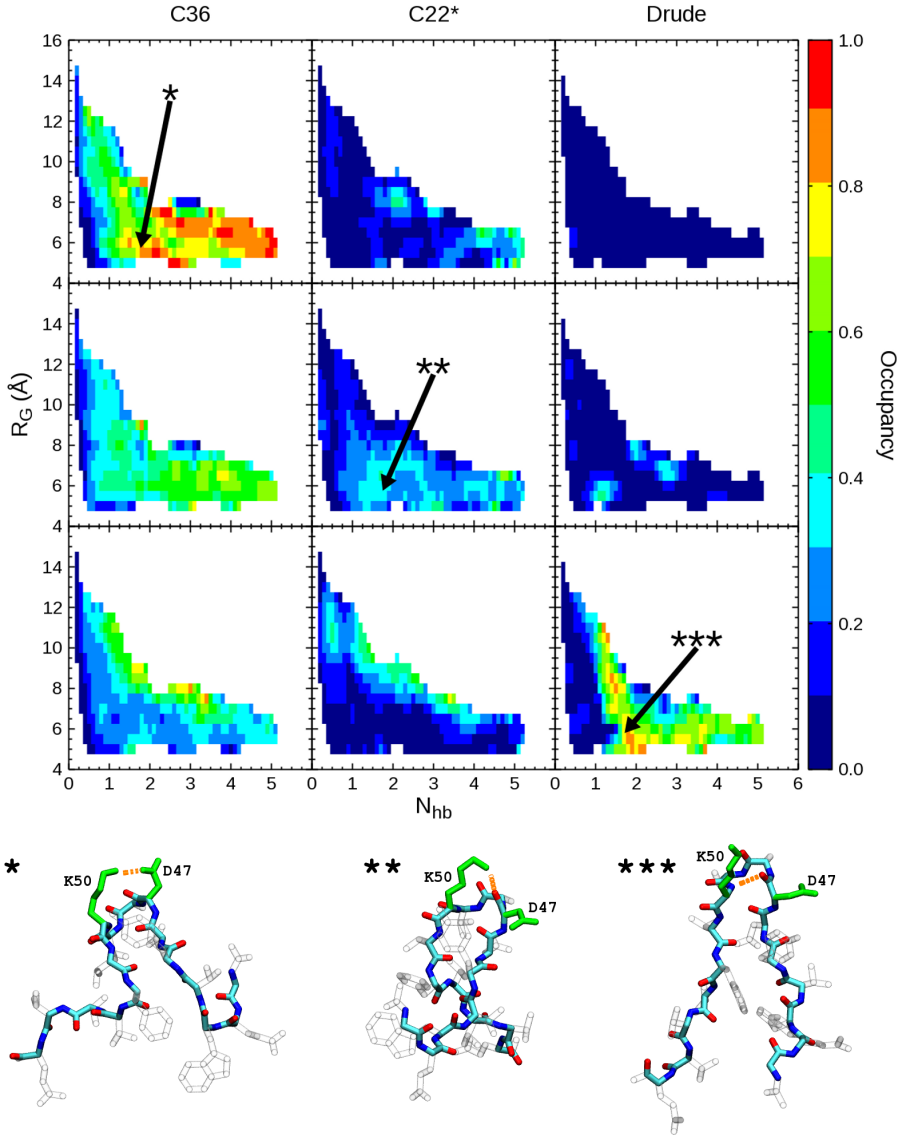


Figure B.10: *Asp47–Lys50 interactions in the turn region of GB1.* Occupation of (top) Asp47–Lys50 salt bridge, (center) Asp47 backbone–Lys50 side chain hydrogen bond, and (bottom) Asp47 backbone–Lys50 backbone hydrogen bond for (left) C36, (middle) C22*, and (right) Drude-2013 REUS trajectories. Salt Bridges and hydrogen bonds were considered occupied when the distance between their respective atoms was <3.4 Å. Representative structures are shown in licorice representation for (*) C36, (**) C22*, and (***) Drude-2013, with their coordinates indicated in the plots. Backbone atoms are colored by atom name. Asp47 and Lys50 side chains are shown in green, while all other side chains are transparent. Hydrogen bonds and salt bridges are shown as orange, dashed lines.

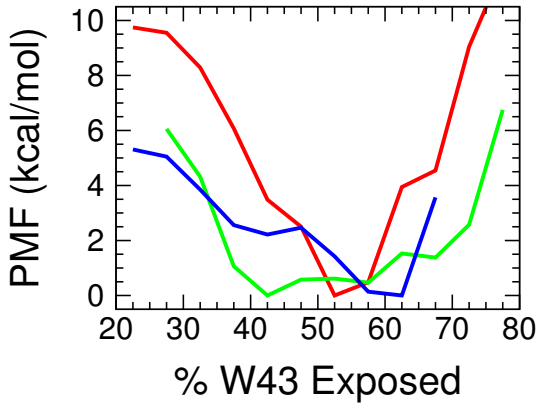


Figure B.11: *1D PMFs of GB1 folding for fraction of exposed Trp43 side chain.* PMFs calculated by integrating 2D REUS PMFs over Trp43 SASA fraction for (red) C36, (green) C22*, and (blue) unmodified Drude.

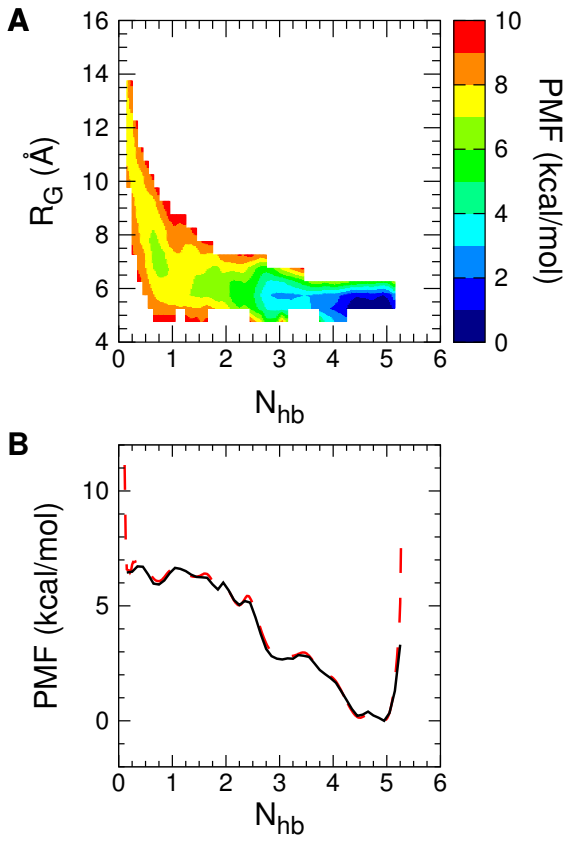


Figure B.12: *PMFs of GB1 folding for C36m using reweighted C36 REUS trajectories.* C36 REUS trajectories were reweighted using C36m parameters according to Eq. 4 from the main text. (A) 2D PMF. (B) 1D PMF along the N_{hb} coordinate for (black, solid line) C36m and (red, dashed line) C36.

REFERENCES

[1] J. R. Perilla, J. A. Hadden, B. C. Goh, C. G. Mayne, and K. Schulten, "All-atom molecular dynamics of virus capsids as drug targets," *J. Phys. Chem. Lett.*, vol. 7, no. 10, pp. 1836–1844, 2016.

[2] M. Sener, J. Strumpfer, A. Singharoy, C. N. Hunter, and K. Schulten, "Overall energy conversion efficiency of a photosynthetic vesicle," *eLIFE*, vol. 5, e09541, 2016.

[3] K. Lindorff-Larsen, P. Maragakis, S. Piana, and D. E. Shaw, "Picosecond to millisecond structural dynamics in human ubiquitin," *J. Phys. Chem. B*, vol. 120, no. 33, pp. 8313–8320, 2016.

[4] D. Guo, A. C. Pan, R. O. Dror, T. Mocking, R. Liu, L. H. Heitman, D. E. Shaw, and A. P. IJzerman, "Molecular basis of ligand dissociation from the adenosine A2A receptor," *Mol. Pharmacol.*, vol. 89, pp. 485–491, 2016.

[5] M. Junker, R. N. Besingi, and P. L. Clark, "Vectorial transport and folding of an autotransporter virulence protein during outer membrane secretion," *Mol. Microbiol.*, vol. 71, no. 5, pp. 1323–1332, 2009.

[6] M. Junker, C. C. Schuster, A. V. McDonnell, K. A. Sorg, M. C. Finn, D. Berger, and P. L. Clark, "Pertactin β-helix folding mechanism suggests common themes for the secretion and folding of autotransporter proteins," *Proc. Natl. Acad. Sci. USA*, vol. 103, no. 13, pp. 4918–4923, 2006.

[7] M. Junker and P. L. Clark, "Slow formation of aggregation-resistant β-sheet folding intermediates," *Proteins: Struct., Func., Bioinf.*, vol. 78, pp. 812–824, 2010.

[8] W. Wojtas-Niziurski, Y. Meng, B. Roux, and S. Bernèche, "Self-learning adaptive umbrella sampling method for the determination of free energy landscapes in multiple dimensions," *J. Chem. Theory Comput.*, vol. 9, pp. 1885–1895, 2013.

[9] *Federal R&D in the FY 2016 budget: An overview*, AAAS, Accessed: 2017-04-02.

[10] *2016 top markets report: Pharmaceuticals*, US Department of Commerce, International Trade Administration, Accessed: 2017-04-02.

[11] X.-Y. Meng, H.-X. Zhang, M. Mezei, and M. Cui, "Molecular docking: A powerful approach for structure-based drug discovery," *Curr. Comput. Aided Drug Des.*, vol. 7, no. 2, pp. 146–157, 2011.

[12] D. Guo, T. Mulder-Krieger, A. P. IJzerman, and L. H. Heitman, "Functional efficacy of adenosine A2A receptor agonists is positively correlated to their receptor residence time," *Br. J. Pharmacol.*, vol. 166, no. 6, pp. 1848–1859, 2012.

[13] "About antimicrobial resistance,"

[14] S. B. Levy and B. Marshall, "Antibacterial resistance worldwide: causes, challenges and responses," *Nat. Med.*, vol. 10, no. 12 Suppl, S122–S129, 2004.

[15] D. Du, H. W. van Veen, and B. F. Luisi, "Assembly and operation of bacterial tripartite multidrug efflux pumps," *Trends Microbiol.*, vol. 23, no. 5, pp. 311–319, 2015.

[16] N. Abdali, J. M. Parks, K. M. Haynes, J. L. Chaney, A. T. Green, D. Wolloscheck, J. K. Walker, V. V. Rybenkov, J. Baudry, J. C. Smith, and H. I. Zgurskaya, "Reviving antibiotics: efflux pump inhibitors that interact with AcrA, a membrane fusion proein of the AcrAB-TolC multridrug efflux pump," *ACS Infect. Dis.*, vol. 3, pp. 89–98, 2017.

[17] K. M. Haynes, N. Abdali, V. Jhawar, H. I. Zgurskaya, J. M. Parks, A. T. Green, J. Baudry, V. V. Rybenkov, J. C. Smith, and J. K. Walker, "Identification and structure-activity relationships of novel compounds that potentiate the activities of antibiotics in *Escherichia coli*," *J. Med. Chem.*, vol. 60, pp. 6205–6219, 2017.

[18] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973.

[19] L. Pauling, R. B. Corey, and H. R. Branson, "The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain," *Proc. Natl. Acad. Sci. USA*, vol. 37, pp. 205–211, 1951.

[20] L. Pauling and R. B. Corey, "The pleated sheet, a new layer configuration of polypeptide chains," *Proc. Natl. Acad. Sci. USA*, vol. 37, pp. 251–256, 1951.

[21] D. J. Barlow and J. M. Thornton, "Helix geometry in proteins," *J. Mol. Biol.*, vol. 201, no. 3, pp. 601–619, 1988.

[22] A. F. Cullis, H. M. Dintzis, and M. F. Perutz, "X-ray analysis of haemoglobin," *Conference on Hemoglobin. Nat. Acad. Sei., National Research Council, Washington, DC, Publication*, no. 557, p. 50, 1958.

[23] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, and H. Wyckoff, "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis," *Nature*, vol. 181, pp. 662–666, 1958.

[24] C. Levinthal, "How to fold graciously," *Mossbaur Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois*, pp. 22–24, 1969.

[25] B. H. Zimm and J. K. Bragg, "Theory of phase transition between helix and random coil in polypeptide chains," *J. Chem. Phys.*, vol. 31, pp. 526–535, 1959.

[26] D. M. Holtzman, J. C. Morris, and A. M. Goate, "Alzheimer's disease: The challenge of the second century," *Sci. Transl. Med.*, vol. 3, 77sr1–77sr1, 2011.

[27] W. Han and K. Schulten, "Fibril elongation by a$\beta 17--42$: Kinetic network analysis of hybrid-resolution molecular dynamics simulations," *J. Am. Chem. Soc.*, vol. 136, pp. 12 450–12 460, 2014.

[28] S. B. Prusiner, M. P. McKinley, K. A. Bowman, D. C. Bolton, P. E. Bendheim, D. F. Groth, and G. G. Glenner, "Scrapie prions aggregate to form amyloid-like birefringent fibers," vol. 35, no. 2, pp. 349–358, 1983.

[29] Y. Shi, Z. Xia, J. Zhang, R. Best, C. Wu, J. W. Ponder, and P Ren, "Polarizable atomic multipole-based AMOEBA force field for proteins," *J. Chem. Theory Comput.*, vol. 9, no. 9, pp. 4046–4063, 2013.

[30] S. Patel and C. L. Brooks, "CHARMM fluctuating charge force field for proteins: I Parameterization and application to bulk organic liquid simulations," *J. Comput. Chem.*, vol. 25, pp. 1–16, 2004.

[31] S Patel, A. D. MacKerrell, and C. L. Brooks, "CHARMM fluctuating charge force field for proteins: II Protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model$journal = jcc, volume = 25, number =, pages =15$

[32] W. Xie, M. Orozco, D. G. Truhlar, and J. Gao, "X-Pol potential: An electronic structure-based force field for molecular dynamics simulation of a solvated protein in water," *J. Chem. Theory Comput.*, vol. 5, pp. 459–467, 2009.

[33] P. E. M. Lopes, J. Huang, J. Shim, Y. Luo, H. Li, B. Roux, and A. D. MacKerell Jr, "Polarizable force field for peptides and proteins based on the classical Drude oscillator," *J. Chem. Theory Comput.*, vol. 9, pp. 5430–5449, 2013.

[34] A. Hazel, C. Chipot, and J. C. Gumbart, "Thermodynamics of deca-alanine folding in water," *J. Chem. Theory Comput.*, vol. 10, pp. 2836–2844, 2014.

[35] J. Gumbart, C. Chipot, and K. Schulten, "Free energy of nascent-chain folding in the translocon," *J. Am. Chem. Soc.*, vol. 133, pp. 7602–7607, 2011.

[36] G. Zhao, J. R. Perilla, E. L. Yufenyuy, X. Meng, B. Chen, J. Ning, J. Ahn, A. M. Gronenborn, K. Schulten, C. Aiken, and P. Zhang, "Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics," *Nature*, vol. 497, pp. 643–646, 2013.

[37] .

[38] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers, "Atomic-level characterization of the structural dynamics of proteins," *Science*, vol. 330, pp. 341–346, 2010.

[39] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, "How fast-folding proteins fold," *Science*, vol. 334, pp. 517–520, 2011.

[40] C. Jarzynski, "Nonequilibrium equality for free energy differences," *Phys. Rev. Lett.*, vol. 78, p. 2690, 1997.

[41] A. Laio and M. Parrinello, "Escape free-energy minima," *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 20, pp. 12 562–12 566, 2002.

[42] J. H
enin and C. Chipot, "Overcoming free energy barriers using unconstrained molecular dynamics simulations," *J. Chem. Phys.*, vol. 121, no. 7, pp. 2904–2914, 2004.

[43] .

[44] E. Darve and A. Pohorille, "Calculating free energies using an average force," *J. Chem. Phys.*, vol. 115, p. 9169, 2001.

[45] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, "The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method," *J. Comput. Chem.*, vol. 13, no. 8, pp. 1011–1021, 1992.

[46] P. G. Bolhuis, C. Dellago, and D. Chandler, "Reaction coordinates of biomolecular isomerization," *Proc. Natl. Acad. Sci. USA*, vol. 97, no. 11, pp. 5877–5882, 2000.

[47] M. Moradi, V. Babin, C. Roland, T. Darden, and C Sagui, "Conformations and free energy landscapes of polyproline peptides," *Proc. Natl. Acad. Sci. USA*, vol. 106, pp. 20 746–20 751, 2009.

[48] M. Moradi and E Tajkhorshid, "Mechanistic picture for conformational transition of a membrane transporter at atomic resolution," *Proc. Natl. Acad. Sci. USA*, vol. 110, pp. 18 916–18 921, 2013.

[49] M. Moradi, V. Babin, C. Roland, and C Sagui, "Reaction path ensemble of B-Z-DNA transition: a comprehensive atomistic study," *Nucleic Acids Res.*, vol. 41, pp. 33–43, 2013.

[50] Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," *Chem. Phys. Lett.*, vol. 314, pp. 141–151, 1999.

[51] Y. Sugita, A. Kitao, and Y. Okamoto, "Multidimensional replica-exchange method for free-energy calculations," *J. Chem. Phys.*, vol. 113, p. 6042, 2000.

[52] E. Darve, D. Rodríguez-Gómez, and A. Pohorille, "Adaptive biasing force method for scalar and vector free energy calculations," *J. Chem. Phys.*, vol. 128, p. 144 120, 2008.

[53] A. Barducci, G. Bussi, and M. Parrinello, "Well-tempered metadynamics: A smoothly converging and tunable free-energy method," *Phys. Rev. Lett.*, vol. 100, p. 020 603, 2008.

[54] G. M. Torrie and J. P. Valleau, "Nonphysical smpaling distributions in Monte Carlo free-energy estimation: umbrella sampling," *J. Chem. Phys.*, vol. 23, pp. 187–199, 1977.

[55] E. Rosta and G. Hummer, "Free energies from dynamic weighted histogram analysis using unbiased Markov state model," *J. Chem. Theory Comput.*, vol. 11, pp. 276–285, 2014.

[56] H. Wua, F. Paula, C. Wehmeyera, and F. Noé, "Multiensemble Markov models of molecular thermodynamics and kinetics," *Proc. Natl. Acad. Sci. USA*, vol. 113, no. 23, E3221–E3230, 2016.

[57] J. Kästner and W. Thiel, "Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method: "Umbrella integration"," *J. Chem. Phys.*, vol. 123, p. 144 104, 2005.

[58] Y. Meng and B. Roux, "Efficient determination of free energy landscapes in multiple dimensions from biased umbrella sampling simulations using linear regression," *J. Chem. Theory Comput.*, vol. 11, pp. 3523–3529, 2015.

[59] C. Zhang, C.-L. Lai, and B. M. Pettitt, "Accelerating the weighted histogram analysis method by direct inversion in the iterative subspace," *Mol. Sim.*, vol. 42, no. 13, pp. 1079–1089, 2016.

[60] S. Park and W. Im, "Theory of adaptive optimization for umbrella sampling," *J. Chem. Theory Comput.*, vol. 10, pp. 2719–2728, 2014.

[61] R. L. Baldwin and G. D. Rose, "Is protein folding hierarchic? ii. folding intermediates and transition states.," *Trends Biochem. Sci.*, vol. 24, pp. 77–83, 1999.

[62] ——, "Is protein folding hierarchic? i. local structure and peptide folding.," *Trends Biochem. Sci.*, vol. 24, pp. 26–33, 1999.

[63] S. Williams, T. P. Causgrove, R. Gilmanshin, K. S. Fang, R. H. Callender, W. H. Woodruff, and R. B. Dyer, "Fast events in protein folding: helix melting and formation in a small peptide," *Biochemistry*, vol. 35, pp. 691–697, 1996.

[64] W. S. Young and C. L. Brooks III, "A microscopic view of helix propagation: N and C-terminal helix growth in alanine helices," *J. Mol. Biol.*, vol. 259, pp. 560–572, 1996.

[65] D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. R. Ho, D. J. Ierardi, I. Kolossváry, J. L. Klepeis, T. Layman, C. McLeavey, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, and S. C. Wang, "Anton, a special-purpose machine for molecular dynamics simulation," *SIGARCH Comput. Archit. News*, vol. 35, pp. 1–12, 2007.

[66] J. L. Klepeis, K. Lindorff-Larsen, R. O. Dror, and D. E. Shaw, "Long-timescale molecular dynamics simulations of protein structure and function," *Curr. Opin. Struct. Biol.*, vol. 19, pp. 120–127, 2009.

[67] G. R. Bowman, V. A. Voelz, and V. S. Pande, "Taming the complexity of protein folding," *Curr. Opin. Struct. Biol.*, vol. 21, pp. 4–11, 2011.

[68] P. L. Freddolino and K. Schulten, "Common structural transitions in explicit-solvent simulations of villin headpiece folding," *Biophys. J.*, vol. 97, pp. 2338–2347, 2009.

[69] G. R. Bowman and V. S. Pande, "Protein folded states are kinetic hubs," *Proc. Natl. Acad. Sci. USA*, vol. 107, pp. 10 890–10 895, 2010.

[70] K. A. Beauchamp, R. McGibbon, Y. S. Lin, and V. S. Pande, "Simple few-state models reveal hidden complexity in protein folding," *Proc. Natl. Acad. Sci. USA*, 2012.

[71] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, "Atomic-level description of ubiquitin folding," *Proc. Natl. Acad. Sci. USA*, vol. 110, pp. 5915–5920, 2013.

[72] C. Chipot and A. Pohorille, Eds., *Free energy calculations. Theory and applications in chemistry and biology*. New York: Springer Verlag, 2007.

[73] T. Lelièvre, G. Stoltz, and M. Rousset, *Free energy computations: A mathematical perspective*. London, UK: Imperial College Press, 2010.

[74] J. Tirado-Rives, D. S. Maxwell, and W. L. Jorgensen, "Molecular dynamics and monte carlo simulations favor the α–helical form for alanine–based peptides in water," *J. Am. Chem. Soc.*, vol. 115, pp. 11 590–11 593, 1993.

[75] C. Chipot and A. Pohorille, "Folding and translocation of the undecamer of poly–L–leucine across the water–hexane interface. a multi–nanosecond molecular dynamics study," *J. Am. Chem. Soc.*, vol. 120, pp. 11 912–11 924, 1998.

[76] O. Collet and C. Chipot, "Non–arrhenius behavior in the unfolding of a short, hydrophobic α–helix. complementarity of molecular dynamics and lattice model simulations," *J. Am. Chem. Soc.*, vol. 125, pp. 6573–6580, 2003.

[77] J. Hénin, K. Schulten, and C. Chipot, "Conformational equilibrium in alanine-rich peptides probed by reversible stretching simulations," *J. Phys. Chem. B*, vol. 110, pp. 16 718–16 723, 2006.

[78] T. J. Gaborek, C. Chipot, and J. D. Madura, "Conformational free-energy landscapes for a peptide in saline environments," *Biophys. J.*, vol. 103, pp. 2513–2520, 2012.

[79] K. O'Neil and W. DeGrado, "A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids," *Science*, vol. 250, no. 4981, pp. 646–651, 1990.

[80] S. Marqusee, V. H. Robbins, and R. L. Baldwin, "Unusually stable helix formation in short alanine-based peptides.," *Proc. Natl. Acad. Sci. USA*, vol. 86, no. 14, pp. 5286–90, 1989.

[81] S. M. Miick, G. V. Martinez, W. R. Fiori, A. P. Todd, and G. L. Millhauser, "Short alanine–based peptides may form 310–helices in aqueous solution," *Nature*, vol. 359, pp. 653–655, 1992.

[82] C. Chipot, "Free energy calculations in biological systems. how useful are they in practice?" In *New algorithms for macromolecular simulation*, B. Leimkuhler, C. Chipot, R. Elber, A. Laaksonen, A. E. Mark, T. Schlick, C. Schütte, and R. Skeel, Eds., vol. 49, Berlin: Springer Verlag, 2005, pp. 183–209.

[83] ——, "Frontiers in free-energy calculations of biological systems," *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, pp. 71–89, 2014.

[84] J. Vreede, J. Juraszek, and P. G. Bolhuis, "Predicting the reaction coordinates of millisecond light-induced conformational changes in photoactive yellow protein," *Proc. Natl. Acad. Sci. USA*, vol. 107, pp. 2397–2402, 2010.

[85] F. Marinelli, F. Pietrucci, A. Laio, and S. Piana, "A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations," *PLoS Comput. Biol.*, vol. 5, e1000452, 2009.

[86] S. Park, F. Khalili-Araghi, E. Tajkhorshid, and K. Schulten, "Free energy calculation from steered molecular dynamics simulations using Jarzynksi's equility," *J. Chem. Phys.*, vol. 119, p. 3559, 2003.

[87] J. Hénin, G. Fiorin, C. Chipot, and M. L. Klein, "Exploring multidimensional free energy landscapes using time-dependent biases on collective variables," *J. Chem. Theory Comput.*, vol. 6, pp. 35–47, 2010.

[88] L. Zheng, M. Chen, and W. Yang, "Random walk in orthogonal space to achieve efficient free-energy simulation of complex systems," *Proc. Natl. Acad. Sci. USA*, vol. 105, no. 51, pp. 20 227–20 232, 2008.

[89] K. Minoukadeh, C. Chipot, and T. Lelièvre, "Potential of mean force calculations: a multiple-walker adaptive biasing force approach," *J. Chem. Theory Comput.*, vol. 6, pp. 1008–1017, 2010.

[90] J. Comer, B. Roux, and C. Chipot, "Achieving ergodic sampling using replica-exchange free-energy calculations," *Mol. Sim.*, vol. 40, 2014.

[91] W. Humphrey, A. Dalke, and K. Schulten, "VMD – Visual Molecular Dynamics," *J. Mol. Graphics*, vol. 14, pp. 33–38, 1996.

[92] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, pp. 926–935, 1983.

[93] A. D. MacKerell, Jr., M. Feig, and C. L. Brooks, III, "Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations," *J. Comput. Chem.*, vol. 25, pp. 1400–1415, 2004.

[94] R. B. Best, J. Mittal, M. Feig, and A. D. MacKerrell Jr, "Inclusion of many-body effects in the additive CHARMM protein CMAP potential results in enhanced cooperativity of α-helix and β-hairpin formation," *Biophys. J.*, vol. 103, pp. 1045–1051, 2012.

[95] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell Jr, "Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ,ψ, and side-chain χ_1 and χ_2 dihedral angles," *J. Chem. Theory Comput.*, vol. 8, pp. 3257–3273, 2012.

[96] S. E. Feller, Y. H. Zhang, R. W. Pastor, and B. R Brooks, "Constant pressure molecular dynamics simulations — The Langevin piston method," *J. Chem. Phys.*, vol. 103, pp. 4613–4621, 1995.

[97] T. A. Darden, D. M. York, and L. G. Pedersen, "Particle mesh Ewald: an N log N method for Ewald sums in large systems," *J. Chem. Phys.*, vol. 98, pp. 10 089–10 092, 1993.

[98] H. C. Andersen, "Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations," *J. Chem. Phys.*, vol. 52, pp. 24–34, 1983.

[99] B. Roux, "Statistical mechanical equilibrium theory of selective ion channels," *Biophys. J.*, vol. 77, pp. 139–153, 1999.

[100] C. Chipot and J. Hénin, "Exploring the free-energy landscape of a short peptide using an average force," *J. Chem. Phys.*, vol. 123, p. 244 906, 2005.

[101] P. Procacci, S. Marsili, A. Barducci, G. F. Signorini, and R. Chelli, "Crooks equation for steered molecular dynamics using a Nosé-Hoover thermostat," *J. Chem. Phys.*, vol. 125, p. 164 101, 2006.

[102] G. Ozer, S. Quirk, and R. Hernandez, "Adaptive steered molecular dynamics: Validation of the selection criterion and benchmarking energetics in vacuum," *J. Chem. Phys.*, vol. 136, p. 215 104, 2012.

[103] R. Chelli, C. Gellini, G. Pietraperzia, E. Giovannelli, and G. Cardini, "Path-breaking schemes for nonequilibrium free energy calculations," *J. Chem. Phys.*, vol. 138, p. 214 109, 2013.

[104] Y. Levy, J. Jortner, and O. M. Becker, "Solvent effects on the energy landscapes and folding kinetics of polyalanine," *Proc. Natl. Acad. Sci. USA*, vol. 98, no. 5, pp. 2188–2193, 2001.

[105] H. Kokubo, C. Y. Hu, and B. M. Pettitt, "Peptide conformational preferences in osmolyte solutions: Transfer free energies of decaalanine," *J. Am. Chem. Soc.*, vol. 133, pp. 1849–1858, 2011.

[106] B. Ensing, A. Laio, M. Parrinello, and M. L. Klein, "A recipe for the computation of the free energy barrier and the lowest free energy path of concerted reactions," *J. Phys. Chem. B*, vol. 109, pp. 6676–6687, 2005.

[107] A. C. Pan, D. Sezer, and B. Roux, "Finding transition pathways using the string method with swarms of trajectories," *J. Phys. Chem. B*, vol. 112, pp. 3432–3440, 2008.

[108] V. Ovchinnikov, M. Karplus, and E. Vanden-Eijnden, "Free energy of conformational transition paths in biomolecules: The string method and its application to myosin VI," *J. Chem. Phys.*, vol. 134, p. 085 103, 2011.

[109] G. Ozer, S. Quirk, and R. Hernandez, "Thermodynamics of decaalanine stretching in water obtained by adaptive steered molecular dynamics simulations," *J. Chem. Theory Comput.*, vol. 8, pp. 4837–4844, 2012.

[110] G. Stirnemann, S. Kang, R. Zhou, and B. J. Berne, "How force unfolding differs from chemical denaturation," *Proc. Natl. Acad. Sci. USA*, vol. 111, pp. 3413–3418, 2014.

[111] C. A. Rohl, A. Chakrabartty, and R. L. Baldwin, "Helix propagation and N-cap propensities of the amino acids measured in alanine-based peptides in 40 volume percent trifluoroethanol," *Protein Science*, vol. 5, pp. 2623–2637, 1996.

[112] S. E. Blondelle, B. Forood, R. A. Houghten, and E. Pérez-Payá, "Polyalanine-based peptides as models for self-associated β-pleated-sheet complexes," *Biochemistry*, vol. 36, pp. 8393–8400, 1997.

[113] L. Williams, K. Kather, and D. S. Kemp, "High helicities of Lys-containing, Ala-rich peptides are primarily attributable to a large, context-dependent Lys stabilization," *J. Am. Chem. Soc.*, vol. 120, pp. 11 033–11 043, 1998.

[114] E. J. Spek, C. A. Olson, Z. Shi, and N. R. Kallenbach, "Alanine is an intrinsic α-helix stabilizing amino acid," *J. Am. Chem. Soc.*, vol. 121, pp. 5571–5572, 1999.

[115] J. Graf, P. H. Nguyen, G. Stock, and H. Schwalbe, "Stucture and dynamics of the homologous series of alanine peptides: A joint molecular dynamics/NMR study," *J. Am. Chem. Soc.*, vol. 129, pp. 1179–1189, 2007.

[116] B. van den Berg, R. J. Ellis, and C. M. Dobson, "Effects of macromolecular crowding on protein folding and aggregation," *EMBO J.*, vol. 18, pp. 6927–6933, 1999.

[117] C. M. Dobson, "Protein folding and misfolding," *Nature*, vol. 426, pp. 884–890, 2003.

[118] R. B. Best, N.-V. Buchete, and G. Hummer, "Are current molecular dynamics force fields too helical?" *Biophys. J.*, vol. 95, pp. L07–L09, 2008.

[119] V. Muñoz and L. Serrano, "Elucidating the folding problem of helical peptides using empirical parameters," *Nat. Struct. Biol.*, vol. 1, pp. 339–409, 1994.

[120] ——, "Elucidating the folding problem of α-helical peptides using empirical parameters, II. Helix macrodipole effects and rational modifications of the helical content of natural peptides," *J. Mol. Biol.*, vol. 245, pp. 275–296, 1994.

[121] ——, "Elucidating the folding problem of α-helical peptides using empirical parameters, III. Temperature and pH dependence," *J. Mol. Biol.*, vol. 245, pp. 297–308, 1994.

[122] ——, "Development of multiple sequence approximation within the Agadir model of α-helix formation. Comparison with Zimm-Bragg and Lifson-Roig formalisms," *Biopolymers*, vol. 41, pp. 495–509, 1997.

[123] E. Lacroix, A. R. Viquera, and L. Serrano, "Elucidating the folding problem of α-helices: Local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters," *J. Mol. Biol.*, vol. 284, pp. 173–191, 1998.

[124] W. Shalongo, L. Dugad, and E. Stellwagen, "Distribution of helicity within the model peptide Acetyl(AAQAA)$_3$amide," *J. Am. Chem. Soc.*, vol. 116, pp. 8288–8293, 1994.

[125] T. D. Romo and A. Grossfield, "Unknown unknowns: the Challenge of systematic and statistical error in molecular dynamics simulations," *Biophys. J.*, vol. 106, pp. 1553–1554, 2014.

[126] K. A. Dill and J. L. MacCallum, "The protein-folding problem, 50 years on," *Science*, vol. 338, pp. 1042–1046, 2012.

[127] R. D. Schaeffer and V. Daggett, "Protein folds and protein folding," *Protein Eng. Des. Sel.*, vol. 24, pp. 11–19, 2011.

[128] F. J. Blanco, G. Rivas, and L. Serrano, "A short linear peptide that folds into a native stable β-hairpin in aqueous solution," *Nat. Struct. Biol.*, vol. 1, pp. 584–590, 1994.

[129] H. Hwang, T. G. McCaslin, A. Hazel, C. V. Pagba, C. M. Nevin, A. Pavlova, B. A. Barry, and J. C. Gumbart, "Redox-driven conformational dynamics in a photosystem-ii-inspired β-hairpin maquette determined through spectroscopy and simulation," *J. Phys. Chem. B*, vol. 121, no. 15, 3536–3545, 2017.

[130] R. M. Fesinmeyer, F. M. Hudson, and N. H. Andersen, "Enhanced hairpin stability through loop design: The case of the protein G B1 domain hairpin," *J. Am. Chem. Soc.*, vol. 126, pp. 7238–7243, 2004.

[131] P. A. Thompson, W. A. Eaton, and J. Hofrichter, "Laser temperature jump study of the helix-coil kinetics of an alanine peptide interpreted with a 'kinetic zipper' model," *Biochemistry*, vol. 36, no. 30, pp. 9200–9210, 1997.

[132] V. Muñoz, E. R. Henry, J. Hofrichter, and W. A. Eaton, "A statistical mechanical model for β-hairpin kinetics," *Proc. Natl. Acad. Sci. USA*, vol. 95, pp. 5872–5879, 1998.

[133] A. R. Dinner, T. Lazaridis, and M. Karplus, "Understanding β-hairpin formation," *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 9068–9073, 1999.

[134] V. S. Pande and D. S. Rokhsar, "Molecular dynamics simulations of unfolding and refolding of a β-hairpin fragment of protein G," *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 9062–9067, 1999.

[135] A. E. García and K. Y. Sanbonmatsu, "Exploring the energy landscape of a β hairpin in explicit solvent," *Proteins: Struct., Func., Bioinf.*, vol. 42, pp. 345–354, 2001.

[136] R. Zhou and B. J. Berne, "Can a continuum solvent model reproduce the free energy landscape of a β-hairpin folding in water?" *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 20, pp. 12 777–12 782, 2002.

[137] Q. Shao, L. Yang, and Y. Q. Gao, "A test of implicit solvent models on the folding simulation of the GB1 peptide," *J. Chem. Phys.*, vol. 130, p. 195 104, 2009.

[138] R. G. Best and J. Mittal, "Free-energy landscape of the GB1 hairpin in all-atom explicit solvent simulations with different force fields: similarites and differences," *Proteins: Struct., Func., Bioinf.*, vol. 79, pp. 1318–1328, 2011.

[139] S. W. Rick and S. J. Stuart, "Potentials and algorithms for incorporating polarizability in computer simulations," in *Reviews in Computational Chemistry*, K. B. Lipkowitz and D. B. Boyd, Eds., vol. 18, Hoboken, NJ: John Wiley & Sons, Inc., 2002.

[140] J. W. Ponder and D. A. Case, "Force fields for protein simulations," *Adv. Protein Chem.*, vol. 66, pp. 27–85, 2003.

[141] A. Warshel, M. Kato, and A. V. Pisliakov, "Polarizable force fields: History, test cases, and prospects," *J. Chem. Theory Comput.*, vol. 3, pp. 2034–2045, 2007.

[142] P. E. Lopes, B. Roux, and A. D. MacKerell Jr., "Molecular modeling and dynamics studies with explicit inclusion of electronic polarizability. Theory and applications," *Theor. Chem. Acc.*, vol. 124, pp. 11–28, 2009.

[143] J. Huang and A. D. MacKerell Jr, "Induction of peptide bond dipoles drives cooperative helix formation in the $(AAQAA)_3$ peptide," *Biophys. J.*, vol. 107, pp. 991–997, 2014.

[144] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, "How robust are protein folding simulations with respect to force field parametrization?" *Biophys. J.*, vol. 100, pp. L47–L49, 2011.

[145] A. M. Gronenborn, D. R. Filpula, N. Z. Essig, A. Achari, M. Whitlow, P. T. Wingfield, and G. M. Clore, "A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G," *Science*, vol. 253, pp. 657–661, 1991.

[146] S. Miyamoto and P. A. Kollman, "Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models," *J. Comput. Chem.*, vol. 13, pp. 952–962, 1992.

[147] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, "Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of *n*-alkanes," *J. Comp. Phys.*, vol. 23, pp. 327–341, 1977.

[148] J. Lee, X. Cheng, J. M. Swails, M. S. Yeom, P. K. Eastman, J. A. Lemkul, S. Wei, J. Buckner, J. C. Jeong, Y. Qi, S. Jo, V. S. Pande, D. A. Case, C. L. Brooks, A. D. MacKerell, J. B. Klauda, and W. Im, "CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field," *J. Chem. Theory Comput.*, vol. 12, pp. 405–413, 2016.

[149] S. Jo, T. Kim, V. G. Iyer, and W. Im, "CHARMM-GUI: a web-based graphical user interface for CHARMM," *J. Comput. Chem.*, vol. 29, pp. 1859–1865, 2008.

[150] Y. Luo, W. Jiang, H. Yu, A. D. MacKerell Jr, and B. Roux, "Simulation study of ion pairing in concentrated aqueous salt solutions with a polarizable force field," *Faraday Discuss.*, vol. 160, pp. 135–224, 2013.

[151] G. Lamoureux, E. Harder, I. V. Vorobyov, B. Roux, and A. D. MacKerell Jr, "A polarizable model of water for molecular dynamics simulations of biomolecules," *Chem. Phys. Lett.*, vol. 418, pp. 245–249, 2006.

[152] B. T. Thole, "Molecular polarizabilities calculated with a modified dipole interaction," *J. Chem. Phys.*, vol. 59, pp. 341–350, 1981.

[153] M. R. Shirts, D. L. Mobley, J. D. Chodera, and V. S. Pande, "Accurate and efficient corrections for missing dispersion interactions in molecular simulations," *J. Phys. Chem. B*, vol. 111, pp. 13 052–13 063, 2007.

[154] G. Fiorin, M. L. Klein, and J. Hénin, "Using collective variables to drive molecular dynamics simulations," *Mol. Phys.*, vol. 111, pp. 3345–3362, 2013.

[155] A. Grossfield, "WHAM: the weighted histogram analysis method, version 2.0.6," URL: http://membrane.urmc.rochester.edu/content/wham.

[156] M. Sotomayor and K. Schulten, "Single-molecule experiments in vitro and in silico," *Science*, vol. 316, pp. 1144–1148, 2007.

[157] B. R. Brooks, C. L. Brooks, A. D. MacKerell Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, "CHARMM: the biomolecular simulation program," *J. Comput. Chem.*, vol. 30, no. 10, pp. 1545–1614, 2009.

[158] Y. Deng and B. Roux, "Hydration of amino acid side chains:nonpolar and electrostatic contributions calculated from staged molecular dynamics free energy simulations with explicit water molecules," *J. Phys. Chem. B*, vol. 108, no. 42, pp. 16 567–16 576, 2004.

[159] ——, "Computations of standard binding free energies with molecular dynamics simulations," *J. Phys. Chem. B*, vol. 113, no. 8, pp. 2234–2246, 2009.

[160] E. Harder and B. Roux, "On the origin of the electrostatic potential difference at a liquid-vacuum interface," *J. Chem. Phys.*, vol. 129, no. 23, p. 234 706, 2008.

[161] H. Yu, T. W. Whitfield, E. Harder, G. Lamoureux, I. Vorobyov, V. M. Anisimov, A. D. MacKerell, and B. Roux, "Simulating monovalent and divalent ions in aqueous solution using a drude polarizable force field," *J. Chem. Theory Comput.*, vol. 6, no. 3, pp. 774–786, 2010.

[162] B. Zagrovic, E. J. Sorin, and V. Pande, "β-Hairpin folding simulations in atomistic detail using an implicit solvent model," *J. Mol. Biol.*, vol. 313, pp. 151–169, 2001.

[163] R. Zhou, B. Berne, and R. Germain, "The fre energy landscape for β hairpin folding in explicit water," *Proc. Natl. Acad. Sci. USA*, vol. 98, no. 26, pp. 14 931–14 936, 2001.

[164] D. A. Evans and D. J. Wales, "Folding of the GB1 hairpin peptide from discrete path sampling," *J. Chem. Phys.*, vol. 121, p. 1080, 2004.

[165] G. Bussi, F. L. Gervasio, A. Laio, and M. Parrinello, "Free-energy landscape for β hairpin folding from combined parallel tempering and metadynamics," *J. Am. Chem. Soc.*, vol. 128, pp. 13 435–13 441, 2006.

[166] T. Yoda, Y. Sugita, and Y. Okamoto, "Cooperative folding mechanism of a β-hairpin peptide studied by a multicanonical replica-exchange molecular dynamics simulation," *Proteins: Struct., Func., Bioinf.*, vol. 66, pp. 846–859, 2007.

[167] A. T. Macias and A. D. MacKerell Jr., "CH/π interactions involving aromatic amino acids: Refinement of the CHARMM tryptophan force field," *J. Comput. Chem.*, vol. 26, no. 14, pp. 1452–1463, 2005.

[168] G. König and S. Boresch, "Hydration free energies of amino acids: why side chain analog data are not enough," *J. Phys. Chem. B*, vol. 113, no. 26, pp. 8967–8974, 2009.

[169] J. L. MacCallum and D. P. Tieleman, "Calculation of the watercyclohexane transfer free energies of neutral amino acid side-chain analogs using the opls all-atom force field," *J. Comput. Chem.*, vol. 24, no. 15, pp. 1930–1935, 2003.

[170] M. R. Shirts and V. S. Pande, "Solvation free energies of amino acid side chain analogs for common molecular mechanics water models," *J. Chem. Phys.*, vol. 122, no. 13, p. 134 508, 2005.

[171] B. Hess and N. F. A. van der Vegt, "Hydration thermodynamic properties of amino acid analogues: a systematic comparison of biomolecular force fields and water models," *J. Phys. Chem. B*, vol. 110, pp. 17 616–17 626, 2006.

[172] D. Sitkoff, K. A. Sharp, and B. Honig, "Accurate calculation of hydration free energies using macroscopic solvent models," *J. Phys. Chem.*, vol. 98, no. 7, pp. 1978–1988, 1994.

[173] C. C. Chambers, G. D. Hawkins, C. J. Cramer, and D. G. Truhlar, "Model for aqueous solvation based on class IV atomic charges and first solvation shell effects," *J. Phys. Chem.*, vol. 100, no. 40, pp. 16 385–16 398, 1996.

[174] Y. Mandel-Gutfreund and L. M. Gregoret, "On the significance of alternating patterns of polar and non-polar residues in beta-strands," *J. Mol. Biol.*, vol. 323, pp. 453–461, 2002.

[175] C. M. Davis and R. B. Dyer, "The role of electrostatic interactions in folding of β-proteins," *J. Am. Chem. Soc.*, vol. 138, no. 4, pp. 1456–1464, 2016.

[176] J. H. Tomlinson, S. Ullah, P. E. Hansen, and M. P. Williamson, "Characterization of salt bridges to lysines in the protein G B1 domain," *J. Am. Chem. Soc.*, vol. 131, pp. 4674–4684, 2009.

[177] M. S. Najor, K. W. Olsen, D. J. Graham, and D. M. de Freitas, "Contribution of each Trp residue toward the intrinsic fluorescence of the G$i\alpha 1$ protein," *Prot. Sci.*, vol. 23, pp. 1392–1402, 2014.

[178] T. Cellmer, M. Buscaglia, E. R. Henry, J. Hofrichter, and W. A. Eaton, "Making connections between ultrafast protein folding kinetics and molecular dynamics simulations," *Proc. Natl. Acad. Sci. USA*, vol. 108, no. 15, pp. 6103–6108, 2011.

[179] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, and T. Head-Gordon, "Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew," *J. Chem. Phys.*, vol. 120, no. 20, pp. 9665–9678, 2004.

[180] J. L. F. Abascal and C. Vega, "A general purpose model for the condensed phases of water: TIP4P/2005," *J. Chem. Phys.*, vol. 123, p. 234 505, 2005.

[181] S. Piana, A. G. Donchev, P. Robustelli, and D. E. Shaw, "Water dispersion interactions strongly influence simulated structural properties of disordered protein states," vol. 119, pp. 5113–5123, 2015.

[182] P. E. Lopes, G. Lamoureux, and A. D. MacKerell Jr., "Polarizable empirical force field for nitrogen-containing heteroaromatic compounds based on the classical Drude oscillator," *J. Comput. Chem.*, vol. 30, pp. 1821–1838, 2009.

[183] L. Li, C. Li, Z. Zhang, and E. Alexov, "On the dielectric "constant" of proteins: Smooth dielectric function for macromolecular modeling and its implementation in DelPhi," *J. Chem. Theory Comput.*, vol. 9, pp. 2126–2136, 2013.

[184] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller, and A. D. MacKerell Jr, "CHARMM36m: an improved force field for folded and intrinsically disordered proteins," *Nat. Methods*, vol. 14, pp. 71–73, 2017.

[185] C. J. Oomen, P. van Ulsen, P. van Gelder, M. Feijen, J. Tommassen, and P. Gros, "Structure of the translocator domain of a bacterial autotransporter," *EMBO J.*, vol. 23, no. 6, pp. 1257–1266, 2004.

[186] T. J. Barnard, N. Dautin, P. Lukacik, H. D. Bernstein, and S. K. Buchanan, "Autotransporter structure reveals intra-barrel cleavage followed by conformational changes," *Nat. Struct. Mol. Biol.*, vol. 14, no. 12, pp. 1214–1220, 2007.

[187] B. van den Berg, "Crystal structure of a full-length autotransporter," *J. Mol. Biol.*, vol. 396, no. 3, pp. 627–633, 2010.

[188] D. L. Leyton, A. E. Rossiter, and I. R. Henderson, "From self sufficiency to dependence: mechanisms and factors important for autotransporter biogenesis," *Nat. Rev. Microbiol.*, vol. 10, pp. 213–225, 2012.

[189] R. Ieva and H. D. Bernstein, "Interaction of an autotransporter passenger domain with bama during its translocation across the bacterial outer membrane," *Proc. Natl. Acad. Sci. USA*, vol. 106, no. 45, pp. 19 120–19 125, 2009.

[190] J. H. Peterson, P. Tian, R. Ieva, N. Dautin, and H. D. Bernstein, "Secretion of a bacterial virulence factor is driven by the folding of a c-terminal segment," *Proc. Natl. Acad. Sci. USA*, vol. 107, no. 41, pp. 17 739–17 744, 2010.

[191] J. P. Renn, M. Junker, R. H. Besingi, E. Braselmann, and P. L. Clark, "ATP-independent control of autotransporter virulence protein transport via the folding properties of the secreted protein," *Chem. Biol.*, vol. 19, pp. 287–296, 2012.

[192] W. Kang'ethe and H. D. Bernstein, "Charge-dependent secretion of an intrinsicially disordered protein via the autotransporter pathway," *Proc. Natl. Acad. Sci. USA*, vol. 110, no. 45, E4246–E4255, 2013.

[193] R. Ieva, P. Tian, J. H. Peterson, and H. D. Bernstein, "Sequential and spatially restricted interactions of assembly factors with an autotransporter β domain," *Proc. Natl. Acad. Sci. USA*, vol. 108, no. 31, E343–E391, 2011.

[194] O. Pavlova, J. H. Peterson, R. Ieva, and H. D. Bernstein, "Mechanistic link between β barrel assembly and the initiation of autotransporter secretion," *Proc. Natl. Acad. Sci. USA*, vol. 110, no. 10, E938–E947, 2013.

[195] I. Drobnak, E. Braselmann, and P. L. Clark, "Multiple driving forces required for efficient secretion of autotransporter virulence proteins," *J. Biol. Chem.*, vol. 290, no. 16, pp. 10 104–10 116, 2015.

[196] J. P. Renn and P. L. Clark, "A conserved stable core structure in the passsenger domain β-helix of autotransporter virulence proteins," *Biopolymers*, vol. 89, no. 5, pp. 420–427, 2008.

[197] X. Yuan, M. D. Johnson, J. Zhang, A. W. Lo, M. A. Schembri, L. C. Wijeyewickrema, R. N. Pike, G. H. M. Huysmans, I. R. Henderson, and D. L. Leyton, "Molecular basis for the folding of β-helical autotransporter passenger domains," vol. 9, p. 1395, 2018.

[198] Y. Zhai, K. Zhang, Y. Zhu, Q. Zhou, J. Lu, I. Black, X. Pang, A. W. Roszak, X. Zhang, N. W. Isaacs, and F. Sun, "Autotransporter passenger domain secretion requires a hydrophobic cavity at the extracellular entrance of the β-domain pore," *Biochem. J.*, vol. 435, pp. 577–587, 2011.

[199] R. N. Besingi, J. L. Chaney, and P. L. Clark, "An alternative outer membrane secretion mechanism for an autotransporter lacking a C-terminal stable core," *Mol. Microbiol.*, vol. 90, no. 5, pp. 1029–1045, 2013.

[200] E. Braselmann and P. L. Clark, "Autotransporters: the cellular environment reshapes a folding mechanism to promote protein transport," *J. Phys. Chem. Lett.*, vol. 3, pp. 1063–1071, 2012.

[201] P. Emsley, I. G. Charles, N. F. Fairweather, and N. W. Isaacs, "Structure of *Bordetella pertussis* virulence factor P.69 pertactin," *Nature*, vol. 381, pp. 90–92, 1996.

[202] A. J. Hazel, E. T. Walters, C. N. Rowley, and J. C. Gumbart, "Folding free energy landscapes of β-sheets with non-polarizable and polarizable CHARMM force fields," *J. Chem. Phys.*, vol. 149, p. 072 317, 2018.

[203] A. W. Bryan Jr, J. L. Starner-Kreinbrink, R. Hosur, P. L. CLark, and B. Berger, "Structure-based prediction reveals capping motics that inhibit β-helix aggregation," *Proc. Natl. Acad. Sci. USA*, vol. 108, no. 27, pp. 11 099–11 104, 2011.

[204] J.-S. Kim, H. Jeong, S. Song, H.-Y. Kim, K. Lee, J. Hyun, and N.-C. Ha, "Structure of the tripartite multidrug efflux pump AcrAB-TolC suggests an alternative assembly mode," *Mol. Cells*, vol. 38, no. 2, pp. 180–186, 2015.

[205] D. Du, Z. Wang, N. R. James, J. E. Voss, E. Klimont, T. Ohene-Agyei, H. Venter, W. Chiu, and B. F. Luisi, "Structure of the AcrAB-TolC multidrug efflux pump," *Nature*, vol. 509, pp. 512–515, 2014.

[206] Z. Wang, G. Fan, C. F. Hryc, J. N. Blaza, I. I. Serysheva, M. F. Schmid, W. Chiu, B. F. Luisi, and D. Du, "An allosteric transport mehcanism for the AcrAB-TolC multidrug efflux pump," *eLife*, vol. 6, e24905, 2017.

[207] H. I. Zgurskaya and H. Nikaido, "Bypassing the periplasm: reconstitution of the AcrAB multidrug efflux pump of *Escherichia coli*," *Proc. Natl. Acad. Sci. USA*, vol. 96, no. 13, pp. 7190–7195, 1999.

[208] K. M. Pos, "Drug transport mechanism of the AcrB efflux pump," *Biochim. Biophys. Acta*, vol. 1794, no. 5, pp. 782–793, 2009.

[209] K. Poole, "Efflux pumps as antimicrobial resistance mechanisms," *Ann. Med.*, vol. 39, pp. 162–176, 2007.

[210] T. J. Opperman and S. T. Ngyuen, "Recent advances toward a molecular mechanism of efflux pump inhibition.," *Front. Microbiol.*, vol. 6, p. 421, 2015.

[211] F. Husain, M. Humbard, and R. Misra, "Interaction between the TolC and AcrA proteins of a multidrug efflux system of *Escherichia coli*," *J. Bacteriol.*, vol. 186, no. 24, pp. 8533–8536, 2004.

[212] C. A. Elkins and H. Nikaido, "Chimeric analysis of AcrA function reveals the importance of its C-terminal domain in its interaction with the AcrB multidrug efflux pump," *J. Bacteriol.*, vol. 185, no. 18, pp. 5349–5356, 2003.

[213] E. B. Tikhonova and H. I. Zgurskaya, "AcrA, AcrB, and TolC of *Escherichia coli* form a stable intermembrane multidrug efflux complex," *J. Biol. Chem.*, vol. 279, no. 31, pp. 32 116–32 124, 2004.

[214] H. I. Zgurskaya and H. Nikaido, "Cross-linked complex between oligomeric periplasmic lipoprotein AcrA and the inner-membrane-associated multidrug efflux pump AcrB from *Escherichia coli*," *J. Bacteriol.*, vol. 182, no. 15, pp. 4264–4267, 2000.

[215] B. Wang, J. Weng, K. Fan, and W. Wang, "Interdomain flexibility and pH-induced conformational changes of AcrA revealed by molecular dynamics simulations," *J. Phys. Chem. B*, vol. 116, pp. 3411–3420, 2012.

[216] H. Jeong, J.-S. Kim, S. Song, H. Shigematsu, T. Yokoyama, J. Hyun, and N.-C. Ha, "Pseudoatomic structure of the tripartite multidrub efflux pump AcrAB-TolC reveals the intermeshing cogwheel-like interaction between AcrA and TolC," *Structure*, vol. 24, pp. 272–276, 2016.

[217] C. W. Hopkins, S. Le Grand, R. C. Walker, and A. E. Roitberg, "Long-time-step molecular dynamics through hydrogen mass repartitioning," *J. Chem. Theory Comput.*, vol. 11, pp. 1864–1874, 2015.

[218] D. Case, R. Betz, D. Cerutti, T. Cheatham, III, T. Darden, R. Duke, T. Giese, H. Gohlke, A. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K. Merz, G. Monard, H. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, D. R. Roe, A. Roitberg, C. Sagui, C. Simmerling, W. Botello-Smith, J. Swails, R. Walker, J. Wang, R. Wolf, X. Wu, L. Xiao, and P. Kollman, *AMBER 2016*. University of California, San Francisco, 2016.

[219] J. B. Klauda, R. M. Venable, J. A. Freites, J. W. O'Connor, D. J. Tobias, C. Mondragon-Ramirez, I. Vorobyov, A. D. MacKerell Jr., and R. W. Pastor, "Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types," *J. Phys. Chem. B*, vol. 114, pp. 7830–7843, 2010.

[220] G. Krishnamoorthy, D. Wolloscheck, J. W. Weeks, C. Croft, V. V. Rybenkov, and H. I. Zgurskaya, "Breaking the permeability barrier of *Escherichia coli* by controlled hyperporination of the outer membrane.,"

[221] A. D. Vogt and E. D. Cera, "Conformational selection is a dominant mechanism of ligand binding," *Biochemistry*, vol. 52, no. 34, pp. 5723–5729, 2013.

[222] J. W. Weeks, L. M. Nickels, A. T. Ntreh, and H. I. Zgurskaya, "Non-equilivalent roles of two periplasmic subunits in the function and assembly of triclosan pump TriABC from *Pseudomonas aeruginosa*," *Mol. Microbiol.*, vol. 98, pp. 343–356, 2015.

[223] A. T. Ntreh, J. W. Weeks, L. M. Nickels, and H. I. Zgurksaya, "Opening the channel: the two functional interfaces of *Pseudomonas aeruginosa* OpmH with the triclosan efflux pump TriABC," *J. Bacteriol.*, vol. 198, no. 23, pp. 3176–3185, 2016.

[224] V. S. Pande, K. Beauchamp, and G. R. Bowman, "Everything you wanted to know about Markov State Models but were afraid to ask," *Methods*, vol. 52, pp. 99–105, 2010.

[225] M. I. Zimmerman and G. R. Bowman, "FAST conformational searches by balancing exploration/exploitation trade-offs," *J. Chem. Theory Comput.*, vol. 11, pp. 5747–5757, 2015.

[226] Y. Naritomi and S. Fuchigami, "Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: the case of domain motions," *J. Chem. Phys.*, vol. 134, p. 065 101, 2011.

[227] C. R. Schwantes and V. S. Pande, "Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9," *J. Chem. Theory Comput.*, vol. 9, pp. 2000–2009, 2013.

[228] N. Ahalawat and J. Mondal, "Assessment and optimization of collective variables for protein conformational landscape: GB1 β-hairpin as a case study," *J. Chem. Phys.*, vol. 149, p. 094 101, 2018.

[229] H. Wu, F. Nüske, F. Paul, S. Klus, P. Koltai, and F. Noé, "Variational Koopman models: slow collective variables and molecular kinetics from short off-equilibrium simulations," *J. Chem. Phys.*, vol. 146, p. 154 104, 2017.

[230] A. A. Chen and A. E. García, "High-resolution reversible folding of hyperstable RNA tetraloppos using molecular dynamics simulations," *Proc. Natl. Acad. Sci. USA*, vol. 110, pp. 16 820–16 825, 2013.

[231] J. A. Lemkul, J. Huang, B. Roux, and A. D. MacKerell Jr., "An empirical polarizable force field based on the classical Drude oscillator model: Development history and recent applications," *Chem. Rev.*, vol. 116, no. 9, pp. 4982–5013, 2016.

[232] H. M. Khan, C. Grauffel, R. Broer, A. D. MacKerell Jr, R. W. A. Havenith, and N. Reuter, "Improving the force field description of tyrosine-choline cation-π interactions: QM investigation of phenol-N(ME)$_4^+$ interactions," *J. Chem. Theory Comput.*, vol. 12, pp. 5585–5595, 2016.

[233] J. Huang, P. E. M. Lopes, B. Roux, and A. D. MacKerell Jr, "Recent advances in polarizable force fields for macromolecules: microsecond simulations of proteins using the classical Drude oscillator model," *J. Phys. Chem. Lett.*, vol. 5, pp. 3144–3150, 2014.

VITA

Anthony Hazel was born in Orlando, FL in 1989. Throughout his formative years, he always possessed an innate talent for mathematics and a striking curiosity of the world around him. He enjoyed solving mathematics and logic puzzles and reading about such other worldly topics as the big bang, black holes, and time travel. In middle and high schools, he expanded his intellectual pursuits to the mathematical beauty of music, studying the trumpet, piano, and guitar.

Upon graduation in 2008, Anthony traveled only slightly north to Florida State University in Tallahassee, FL, where he initially intended to study music performance and composition. After missing his audition, he chose instead to study engineering, before changing course once again to major in physics and applied mathematics, while also studying chemistry and biophysics. Working under Prof. Peng Xiong in the Physics department, he completed an undergraduate honors thesis characterizing the properties of a novel field effect transistor developed by one of Prof. Xiong's senior grad students, Kan-Sheng Chen. It was during this time, while also attending advanced chemistry labs, that Anthony decided experiments were not for him, and he vowed to never venture too far from the warm glow of his computer screen and the vibrant colors of the chalk board.

In 2012, Anthony continued his journey north to Atlanta, GA, and the Georgia Institute of Technology to pursue his Ph. D. in physics. He quickly joined the lab of Prof. J. C. Gumbart in 2013, where he discovered the wonderful world of biomolecular simulations.